NATIONAL SCIENCE FOUNDATION

MOSAIC

VOLUME 13 NUMBER 1 JANUARY/FEBRUARY 1982

LD S & T C
NEWARK PUBLIC LIBRARY
FILE ROOM

- Shrinking sun
- African lakes
- VLSI technology
- Flexible gender
- Nitrous oxide

Marking the sun's size

In this *Mosaic*

Ever since they were created by the tectonic forces tearing East Africa away from the rest of the continent, the beds of African rift lakes have been accumulating sediment at the rate of some one or two millimeters a year. That's ten to twenty centimeters in a century, up to two meters in a thousand years. And with no oxygen in those depths to decompose organic matter or to support life forms that might disturb the record, a unique environmental history is encoded in those sediments. Events many thousands of years old and only weeks apart in time, says paleoecologist Dan Livingstone of Duke University, may be distinguished from one another.

Livingstone has been interested in those sediments, and the environmental record they hold, for more than a quarter-century. Twenty years ago, from lake-bottom cores 30 meters long, he was able to read the record of ice-age climate in Africa.

In recent years (see "Biology, Behavior, and Environment in Human Evolution," *Mosaic*, Volume 10, Number 2), Livingstone, Neil Opdyke, and others have been more and more driven to go deeper, further back in time. A thousand-meter core, they proposed, would unfold half a million years and important events in a critical span of the maturation of the genus *Homo*. The technologies developed in a decade or more of deep-ocean drilling should make even cores 3,000 meters deep, under 1,000 meters or more of water, a reasonable goal.

In 1981, Livingstone got the chance at least to demonstrate the feasibility of his ideas.

The African rift lakes lie along a continental extension of the tectonic spreading centers that girdle the globe. There are not many places on earth where evidence of the processes that create and destroy the earth's crust are so near at hand. This makes the lake bottoms an important geophysical as well as paleobiological resource. And, as William J. Cromie, who was along on the expedition, reports in this *Mosaic* ("Beneath the African Rift Lakes"), Livingstone and Opdyke joined forces with Duke colleague Bruce Rosendahl, a geophysicist. For reasons that soon begin to appear obvious, representatives of two oil companies—one French and one from the United States—went along.

At least two important things transpired: Livingstone, Rosendahl, and their colleagues showed the feasibility of their proposal, and strong evidence of potentially important petroleum reserves were detected beneath Lakes Tanganyika and Malawi (a first for a basic-research exercise, Rosendahl suggests).

Nobody's troubles are over yet, of course. To pursue their scientific goals, Livingstone, Rosendahl *et al.* must somehow overcome enormous logistical obstacles to transporting the equivalent of a deep-sea drilling operation to within a few kilometers of where Stanley found an earlier Livingstone. And those companies (there are more than two now) seeking to explore further the intriguing evidence of rift-valley petroleum face similar logistical problems, as well as some geopolitical ones. Oil is a resource in which every nation that borders the lakes—and many that don't—will be sure to have an interest.

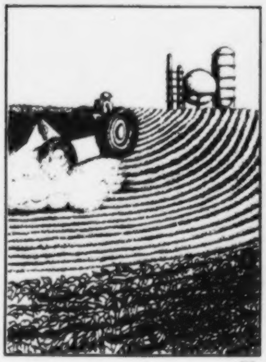
Those problems are interesting, just as the unfolding solutions will be. But equally interesting, on another level entirely, is the message these events appear to convey about the nature of scientific investigation. Perhaps the message has to do with serendipity, perhaps with the intimate relationship between fundamental and applied research. Whichever, it certainly seems to make this point: Virtually no area of scientific research is so esoteric that it might not hold important—if not always so immediate—consequences for what nonscientists sometimes confusedly call the real world.

Warren Kornberg

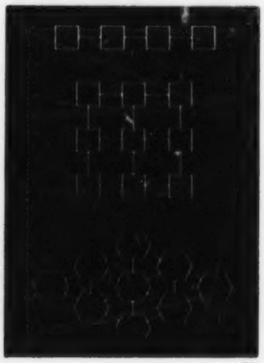
MOSAIC

VOL. 13, NO. 1 JAN./FEB. 1982

2



20



26

Beneath the African Lakes *by William J. Cromie* 2
The ancient lakes of eastern Africa, among the oldest and deepest in the world, contain a record millions of years old—and more than a trace of oil.

The Shrinking Sun *by Diane Johnson* 7
Astronomers searching two centuries of historical records have found evidence that the sun's diameter is changing. The rate of change remains a controversy.

The Gender Changers *by Ben Patrusky* 13
Many lower-order plants and animals are hermaphroditic. The evolutionary and environmental conditions that trigger sexual elasticity are becoming clear.

The Problem of Nitrous Oxide *by Sam Iker* 20
The increase of nitrous oxide in the atmosphere threatens the earth's protective ozone layer. The challenge is to find where it arises and how it can be controlled.

Faster Switches, Smaller Wires, Larger Chips *by Mort La Brecque* 26
Very large scale integrated (VLSI) circuit technology
is revolutionizing the design of computer microcircuits.
Some of the mystery is coming out of the process.

Letters, etc. 33

Trends in Science and Engineering Degrees 34

NATIONAL SCIENCE FOUNDATION

Lewis M. Branscomb, *Chairman*
National Science Board

John B. Slaughter, *Director*
National Science Foundation

Thomas Ubois, *Acting Director*
Office of Government and
Public Programs

Warren Kornberg, Editor
David Heffernan,
Assistant to the Editor
Joyce Latham, Copy Editor

Mosaic is published six times yearly as a source of information for the general public and the scientific and educational communities served by the National Science Foundation. The Director of the National Science Foundation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this agency. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1984. Subscriptions are available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription is $12.00 per year in the United States and possessions and $15.00 all other countries. Single copies are $3.25 domestic and $4.10 foreign. Send check or money order to the U.S. Government Printing Office. Communications other than those concerning paid subscriptions should be addressed to: Editor, *Mosaic*, National Science Foundation, Washington, D.C. 20550. Telephone: 202-357-9776.

Cover. The 1980 solar eclipse over India gave researchers a chance to refine their techniques for accurate determination of the eclipse shadow and thereby the sun's size.

NCAR

Beneath the African Lakes

by William J. Cromie

The ancient lakes of [illegible] among the oldest and [illegible] contain a record [illegible] and [illegible]

ZAIRE
EQUATOR

Many of the best records of the earth's biological and geological history lie at the bottom of oceans and lakes. Cores from the deep-ocean floor contain fossils and undisturbed sediments tens of millions of years old, and much of what is known of the earth's history comes from them. But because seabed sediments accumulate so slowly and because they are reworked by chemical and biological agents, events separated by less than 5,000 years cannot be resolved in them.

That is a gap lakes can fill. With waters that are perennially stratified and bottom sediments that are poor in oxygen and not reworked after deposition, many lakes preserve the detail seabed cores lack.

The common, glacial lakes of the temperate zone are seldom older than 15,000 years. To get longer detailed records, Duke University biologist Daniel A. Livingstone and others proposed drilling beneath the ancient lakes of eastern Africa. The lakes of the Great Rift Valley there, among the oldest and deepest in the world, should contain sedimentary records with a resolution so high that separate seasons can be identified as far back as five million, even ten million, years. Further, techniques developed in the past decade by the ocean-going Deep Sea Drilling Project, managed by the Scripps Institution of Oceanography, put the bottoms of these lakes within reach of coring technology.

Lake-bottom sediments in Africa offer additional information bonuses. Some of the lakes shores were the sites of the earliest known human activities. They contain a record of the climate to which humans and prehumans had to adapt and of the vegetation available for food and their prey. The lakes also form part of a worldwide tectonic rift system, the continental portions of which are little known. And where tectonic and sedimentary activity come together, the possibility of oil always exists. In fact, a recent discovery of what appears to be a major oil reserve under Lake Tanganyika could produce political and economic impacts as profound as the hoped-for scientific results.

Rift valley. East African lakes along the Great Rift Valley contain high-resolution sedimentary records of the earth's biological and geological history, going back millions of years. The *R/V Uganisha* (inset) was used for the survey of Lake Tanganyika.

William J. Cromie

Human and tectonic evolution

In 1978 Livingstone attended a Washington, D.C., meeting of paleoanthropologists considering the best directions to take in researching human origins. Fossils unearthed in the African Rift Valley during previous ten years had revealed that at least five early human and humanlike species walked upright there, the earliest of them going back some 3.5 million years. But the relationships of the species to one another and to modern humans remain in bitter dispute, and knowledge about the surroundings in which they lived and evolved is sparse.

"A detailed sedimentary record from an early hominid homeland should give us a clearer picture of the environment in which humans originated," Livingstone proposed. "It opens the possibility of understanding the impact of climate and vegetation on the development of erect posture, hunting patterns, social grouping, and family life"—in short, on the making of *Homo sapiens*. "The lake drilling idea immediately excited a great deal of interest and enthusiasm," Livingstone recalls. And the excitement spread from anthropologists to other scientists.

The great lakes of Africa lie in a rift valley that splits the continent from the Red Sea to Mozambique, 4,500 kilometers south. Highlands rising 1,500 meters enclose the chain of lakes and provide a steady source of pollen and sediments that reflect the changes in climate and geology. The valley forms a segment of a rift system that splits the earth's crust into some 20 moving section or plates. The valley walls move apart at a rate of about one millimeter a year, apparently tearing East Africa away from the rest of the continent.

Geologists theorize that the rift valley will widen into a waterway like the Red Sea, then into an ocean, recapitulating the breakup of ancient Gondwana into Africa, South America, Antarctica, and Australia. Studying the early stages of such a process on a continent could result in unsurpassed insight into the mechanisms of continental drift. "The lakes in the East African Rift give the best opportunity to see how fast and in what sequence things occur as a continent is pulled apart," notes geologist James Natland of the Deep Sea Drilling Project.

The lake sediments have a high organic content, parts of the subbottom are characterized by high heat flow, and faults slice the lake bed into elevated blocks and depressions. These are all ingredients for the genesis of oil and gas. Indeed, petroleum seeps occur on the shores of Lakes Tanganyika and Mobutu Sese Seko (formerly Lake Albert). If geologists knew more about these lakes, they probably would have a better understanding of the origin and likely locations of petroleum deposits. This is the prospect that led two oil companies—Mobil Exploration and Producing Services and France's government-run Société Nationale Elf Aquitaine—to participate in the African Lakes Drilling Project and the early deliberations about which lakes to test.

Choosing a lake

In August 1980, specialists in paleoanthropology, geology and geophysics, paleoclimatology, and deep-water drilling met in Washington, D.C., to consider candidate lakes. The object was a preliminary survey through which the best lake would be selected.

One obvious candidate: 93-kilometer-long Lake Turkana (formerly Lake Rudolf) in northern Kenya. Turkana was a center of human activity two million to three million years ago and is one of the richest repositories of early hominid fossils in the world. Turkana also offered the promise of excellent local cooperation from the Kenya National Museum, directed by paleoanthropologist Richard Leakey.

"The sedimentary sequence at the bottom of Lake Turkana extends back at least four million years and presents an opportunity to correlate human evolution with changes in climate and vegetation," Leakey notes. "The sediments...should tell us something about how hominids responded to the replacement of dense forests in the Miocene period [twenty-five million to five million years ago] by open woodlands and savanna in the Pliocene [five million to two million years ago]."

On the other hand, "one of the most important items in any core is pollen from wind-pollinated forest species," says Daniel Livingstone. Unfortunately the Turkana area is sparsely vegetated and lacks a good pollen record. Livingstone notes that a combination of pollen, grass cuticles, and diatoms (single-celled lake plants) in cores gives the best picture of past moisture levels, vegetation, and temperature. In addition, fierce winds sometimes sweep Lake Turkana, and it is without docks or shipyard facilities.

Lake Tanganyika, between Tanzania and Zaire, received a high place on the list of possible sites. The world's longest (650 kilometers) and second-deepest (1,500 meters) freshwater lake, it has accumulated wind-blown and river-borne debris from the surrounding mountains and forests for an estimated ten million years. It is accessible by commercial airline and boasts good dock and shipyard facilities. It is a stormy lake, however, and the scientists were concerned about possible interference from some of the bordering nations. Further, Burton Bay, a geologically interesting enclave on Lake Tanganyika's west coast, is reportedly a refuge for smugglers and pirates.

Lake Malawi (once Nyasa), bordering Malawi, Mozambique, and Tanzania, was a possible site because of its position at the southern end of the African Rift Valley. Its advantages include accessibility, a cooperative government, and good dockyard accommodations. On the negative side, its sediments are of unknown age and content, and the lake is large and stormy.

The planners considered Lake Rukwa in Tanzania. Parts of it are periodically dry, and this make it possible to core a sequence of sediments more than two million years old from a dry-land or shallow-water base. The ease of doing this is offset by a variety of local problems, including lack of logistical facilities, severe fuel scarcity, and limited accessibility.

The selection group also discussed Lake Kivu, bordering Zaire and Rwanda, and Lake Mobutu Sese Seko (Albert), flanked by Uganda and Zaire. Lack of accessibility, logistical facilities, and questionable local cooperation make these lakes less appealing than others. An additional disadvantage of Mobutu Sese Seko was listed as "possible piracy."

A hard start

Livingstone, along with colleagues Bruce Rosendahl of Duke and Neil Opdyke of Columbia University, obtained support from the National Science Foundation and supplemental support from Mobil Exploration and Producing Services and Elf Aquitaine. Plans were launched for a geophysical survey of four lakes. Lakes Tanganyika and Malawi received top priority because of their age, sedimentary sequences, and accessibility. Rukwa made the list because of its high tectonic significance and proximity to the other two lakes. Turkana also would be surveyed if time permitted.

Livingstone and Rosendahl, a geophysicist, planned to start work on 700-meter-deep Lake Malawi by April 20, 1981, but a series of mishaps occurred. The ocean-going ship carrying the expedition's equipment sprang a leak, and its cargo had to be loaded onto another vessel. Then the shipment was delayed by customs formalities and a large backlog of cargo scheduled to be carried by rail from Beira in Mozambique to Blantyre in Malawi. In June, the equipment finally arrived by truck at Nkhata Bay on the west coast of Lake Malawi. Researchers from the United States, led by Livingstone and Rosendahl and working with a team of French sedimentologists, loaded the equipment onto the *Orion*, a United Nations fisheries research vessel. The survey finally began on June 10.

Cromie took part in the survey of Lake Tanganyika.

Difficulties did not end there. Winds constantly churned the surface of the 570-kilometer-long lake, and seasickness plagued those aboard the cramped, 15-meter-long vessel. One researcher contracted malaria. Two others were injured, though not seriously, in an automobile accident.

Despite these problems, the expedition collected seismic reflection data for 2,000 kilometers of lake bottom, plus a core 11 meters long. The core contains a record of the early Holocene, 10,000 to 6,000 years ago. This was a time of moist climate when forests were more widespread than they are now. Preservation of diatoms and pollen grains in the sample is excellent, Livingstone reports. "Microfossils in longer cores, extending beyond 12,000 years, should clearly show...alterations in vegetative cover that must have affected the lives of humans and other animals," he says.

Seismic coverage extended from Karonga Bay at the north end of the lake to Monkey Bay at the southern extremity. The heaviest concentration of data covers the area east of Nkhata Bay, halfway along the length of the lake. In the absence of a permit from the government of Mozambique, the southeastern parts of the lake claimed by that nation were not surveyed.

Broken bottom

A major fault system runs along both sides of Lake Malawi and forms the basin that holds the lake. Other faults divide the lake bottom into long, narrow grabens, or depressions, and horsts, or elevations. These involve vertical displacements of as much as a kilometer. The deepest depressions lie under the west side. Going east, the basement rock under the sediments becomes progressively more shallow, via step faults, and is exposed in the region of Likoma Island, off the Mozambique coast. "The degree, intensity, and complexity of deformation beneath Lake Malawi are nothing short of awesome," declares Rosendahl. "Folding and faulting are present throughout the entire lake bed."

Faults that crisscross the elevations and depressions also have vertical displacements, or throws, on the order of tens to hundreds of meters. These faults cut through intensively folded sediments. However, they seldom disrupt the topmost sedimentary layers, which remain almost horizontal, blanket the entire lake, and average 30 meters in thickness. Rosendahl and Livingstone propose that the amount of time represented by this upper sedimentary unit ranges from 30,000 years to 100,000 years.

The Duke researchers believe that the undeformed top layers rest on an unconformity, a gap in the geological record, created when the lake was almost dry and subjected to a long period of erosion. "Most of Lake Malawi as we presently know it is less than 100,000 years old," states Rosendahl.

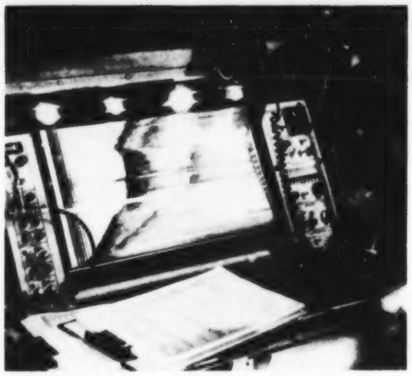
This proposed uplifting and erosion would have killed most of the resident fish. Nevertheless, Lake Malawi contains hundreds of species found nowhere else. Could they have evolved in 100,000 years? "Six species of freshwater fish of the cichlid family are known to have evolved in Lake Nabugabo, Uganda, within 4,000 years," says Livingstone. "At such a rate, 100,000 years might be enough for the evolution of the modern fauna of Lake Malawi, particularly if speciation is punctuated evolution rather than a steady process." (See "The Evolution of Evolution," *Mosaic*, Volume 12, Number 5.) Alternatively, the parent stocks could have survived in a basin that did not dry up.

"The border-fault system undoubtedly represents an extension of the East African Rift into Malawi," Rosendahl says. Hundreds of earthquakes with magnitudes of four or less on the Richter scale attest that this part of the rift is active. "But," notes the Duke geophysicist, "the seismic data reveal no deep features of the kind associated with a Red Sea or Gulf of California-type spreading center, and no systematic thinning of sediments toward any central lake feature. If Lake Malawi represents a segment of an actively rifting plate boundary, it must be in a very juvenile state of development." (See "A Continent Comes Unglued," *Mosaic*, Volume 8, Number 6.)

French geologists believe the rift is very young and predict that a body of water like the Red Sea will open up along it. Continued spreading, theorizes Jacques Le Fournier, leader of the Elf Aquitaine team, will repeat the sequence that formed the South Atlantic Ocean.

Oil discovery

Le Fournier sees similarities between sediments off the coasts of Gabon and Congo



On the ship. Crew members install a compressor aboard the *R/V Uganisha* (top). Jacques Le Fournier (striped shirt, left) and assistants drop pipe for a drilling operation. Bruce Rosendahl examines a geophysical subbottom record in the ship's laboratory.

William J. Cromie

and those in rift valley lakes. The former contain commercial quantities of oil and gas and explain Elf Aquitaine's interest in the latter.

This interest paid dividends—though to whom is still a question—when what appears to be a major oil find was made in Lake Tanganyika. Geophysical soundings reveal a deep basin containing 3,000 meters of sediments located mainly in Zairean waters north of Kalemie, Zaire, and west of Kigoma, Tanzania. Rosendahl believes that the sediments hold "a very large reserve of oil," as well as an environmental record of Africa going back "a minimum of 15 to 20 million years and possibly as far back as 65 to 70 million years."

Besides structural evidence in the seismic soundings, the French team found oil slicks on the lake surface and shallow, oil-soaked sediments in the area. These strong indications increase the probability that oil and natural gas exist in other rift lakes. Indeed, Rosendahl informed the Malawi government that "Lake Malawi possesses a far greater hydrocarbon potential than previously realized." If large amounts of petroleum are there, that will drastically change the political and economic situation in East Africa, with worldwide reverberations.

The oil formed from the bodies of floating and weakly swimming plants and animals (plankton) which rained down on the subsiding grabens of Lakes Tanganyika and Malawi for millions of years. Lack of oxygen below 100 meters preserves the remains and results in sediments with a carbon content as high as 15 to 20 percent. Rosendahl and Le Fournier believe that enough heat to transform this biogenic debris into hydrocarbons rises along the major faults. Shallower faults provide the structural discontinuities that trap pools of oil and gas. The final ingredient in the recipe for hydrocarbons involves reservoirs to hold the oil and gas, and these originate from subaqueous deltas and accumulations of sediments built up by slumping and turbidity currents.

Lake Tanganyika

The primary purpose of the Malawi survey was to evaluate the lake as a potential coring site. Seismic records show "at least 2,000 meters of sediment, an amount consistent with an age of 5 to 10 million years or more," Livingstone says. But the work on Lake Tanganyika suggests better locations because it holds basins that seem to have been continuously filled. No evidence exists of a lakewide unconformity, as is the case for Malawi, and sediment accumulations indicate an age of at least ten million years.

The most striking finding in Lake Tanganyika, as in Malawi, concerns the intensity of faulting and folding. "You can go in any direction at the northern end of the lake and typically find a major fault within a kilometer," notes Rosendahl. An earthquake that sent expedition members scurrying out of a hotel on the lake's edge dramatically showed that the faults remain active.

The big lake, like its neighbor to the south, is defined by major border faults assumed to be caused by tension. But the intense folding of sediments hints that compression is far more prevalent than had been thought. The compression may be associated with shearing at high angles to the border-fault system, Rosendahl suggests. A kink in the centerline of Lake Tanganyika in the vicinity of Kalemie, Zaire, might result from such shear faulting. If this is not the correct interpretation, then geophysicists need to revamp their ideas about folding in tensional regimes, Rosendahl declares.

Tanganyika's fractured horsts and grabens, unlike those in lake Malawi, cut across the border faults and continue on land. Says Rosendahl: "If Tanganyika and Malawi are used as analogs of a continent breaking apart, then continental rifting is far more complex than anyone has modeled to date."

Two models

Geologists have two favorite models of the events in the African Rift zone. One postulates the existence of a hot spot of rising magma in the mantle, causing massive crustal uplifts. (See "The Earth's Fountains of Fire," *Mosaic*, Volume 11, Number 3.) In the early Miocene, about 20 million years ago, the swelling, thinning crust created highlands of a thousand meters and more in East Africa. Magma continued to rise and spread eastward and westward until the blisters cracked open along faults that marked the birth of the Great Rift Valley.

The second model attributes the tension to the subduction, or sinking, under the Iranian plate of the crustal plate holding the Arabian peninsula. Tension at the trailing edge of the Arabian plate pulled open the Red Sea and is tearing Africa apart, according to this scenario.

Some proponents of each model believe that the rift has progressed from north to south and that Lake Malawi is the youngest of the rift lakes. Le Fournier says that an ancient mass of very dense rock (a shield), underlying the Lake Victoria area in southern Kenya and northern Tanzania, resisted faulting and caused the rift to produce an east and west branch. Lakes Mobutu Sese Seko, Edward, Kivu, and Tanganyika lie on the west branch. All along the rift, grabens in the lake beds continue to subside due to the separation of East from Central Africa and also due to the weight of sediments eroded from uplifted mountains and volcanoes. Such features at the edge of the parting plates will, in the French view, become continental shelves in the future.

The geophysical survey's main tools were air guns that fire high-pressure pneumatic blasts into the water every few seconds. The explosions generate sound waves that penetrate the subbottom and are reflected back to the surface from the borders between sedimentary layers and from other structures. Returning echoes are received by hydrophones towed behind the vessel.

Problems in Burundi

When the American team members arrived in Bujumbura, the capital of Burundi, on July 1, they found themselves without a ship or the compressor needed to operate the air guns for the survey of Lake Tanganyika. The compressor had been lost in shipment by air from Blantyre, Malawi. The ship was in Kigoma, Tanzania's northernmost lakeport, so Rosendahl and electronics engineer William Ross Dunseath drove there in a rented Volkswagen.

The two-year-old ship had been on the lake only twice for short periods. The vessel had no lifeboat, working lights, radio, or radar. It carried an anchor but without enough chain to reach the bottom or a windlass powerful enough to raise it. The crew consisted of a Zairean captain who could not read a compass or see well at night and a Tanzanian engineer with limited experience.

Rosendahl and Dunseath rigged lights and lashed the Volkswagen on deck. Following a day of perplexing negotiations with customs officials, they sailed for Bujumbura. After a brief fire in the engine room and a couple of 360-degree turns by the confused captain, Rosendahl got the ship headed upcoast. They lacked charts of the coast or sailing directions, but by following the lights of fishing boats along the shore they managed to reach Bujumbura in 15 hours.

The expedition could not go to sea until the ship's electrical system was completely revamped and the compressor located. Le Fournier suspected the compressor was in Johannesburg, South Africa. However, Burundi has no diplomatic relations with that country and telegrams of inquiry had to go through Belgium. The French geologist decided to fly to Johannesburg. En route, checking a warehouse on a hunch, he found the compressor in Nairobi. No cargo flights were available to Bujumbura, so he air freighted the machine to Kilgali, Rwanda, 300 kilometers away by road. Members of the French team then had to deal with visa problems but managed to rent a truck and retrieve the compressor. They started back to Bujumbura only to find that the Rwanda-Burundi border had been closed for the day, forcing them to lay over until morning.

Success

Christened R/V *Uganisha*, a Swahili word that translates roughly as "patched together," the research vessel finally sailed onto Lake Tanganyika from Bujumbura on August 6. The expedition zigzagged across the central part of the lake, stopping in Kalemie, Zaire, to pick up part of the French team that had surveyed close to shore in a five-meter-long, inflatable rubber boat powered by an outboard engine. On a second trip, the crew performed a detailed geophysical survey and took cores in the northern end of the lake, particularly in the vicinity of the Ruzizi River delta. A final voyage to cover the southern end of the lake had to be cancelled when the air-compressor engine broke down and could not be repaired locally.

By late July, it was evident that Lakes Rukwa and Turkana would be dropped from the survey schedule because the expedition was running out of time and money. Rosendahl hopes to return to Africa in 1982 to finish work on Lake Tanganyika and to conduct geophysical surveys of Turkana and other lakes.

Delays and travail aside, the geophysical surveys of Lakes Malawi and Tanganyika were highly successful. "It was fascinating to work in the ship's laboratory where, night after night, you clearly saw parts of the earth no person has ever seen before," commented Zairean geophysicist Mvumbi Mbungu-Makazu. "We learned more about the geophysics of the East African Rift in three weeks than was learned in the preceding 20 years," adds Rosendahl. He is negotiating with several oil companies, including Elf Aquitaine, to purchase a vessel and perform the surveys of Tanganyika and other lakes. "It would be reasonable for oil companies to join basic scientists in this work," Rosendahl comments.

Livingstone also is satisfied with what was accomplished. Logistics, he has decided, rule out Lakes Kivu, Rukwa, and Mobutu Sese Seko as coring sites. Turkana awaits further exploration. Lake Malawi's suspected unconformity precludes it unless commercial interests decide to work there. "Then we might piggyback the scientific effort onto the exploration for hydrocarbons, which would allow us to obtain the kind of information we want at a cut-rate cost," he says.

Livingstone also is considering Lake Bosumtwi in Ghana because, "pollen and grass cuticle preservation is excellent in this lake, and it is well situated to detect the expansion and contraction of the Sahara. There's no doubt that we'd get a good climate profile from the cores."

First priority for a rift lake goes to Lake Tanganyika, and Livingstone is moving on it as fast as he can. "We will go right to work on the design of a core pipe and a drilling barge," he declared. "That should take a year. Then another year will be needed to construct a barge and ship it to Africa for assembly, probably in Bujumbura."

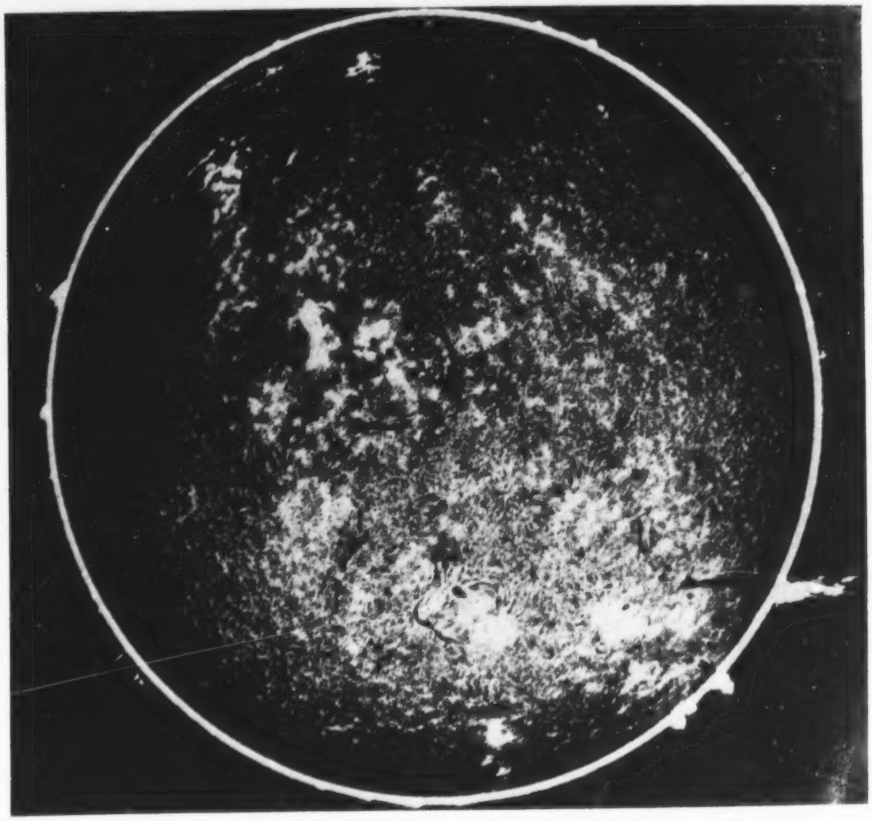
By the time that is done, additional geophysical surveys will have pinpointed a number of desirable coring sites, and data from the African lakes will begin to join those from the Deep Sea Drilling Project, ice cores, and other sources in the effort to put together a picture of the earth's climatic history. In Africa, there will be the additional bonus of information on the response of our ancestors to the environmental changes they encountered. •

The National Science Foundation contributed to the support of the work discussed in this article through its Anthropology Program.

The Shrinking Sun

by Diane Johnson

Astronomers searching two centuries of historical records have found evidence that the sun's diameter is changing. The rate of change remains a controversy.



A composite photograph. Two superimposed pictures taken at different wavelengths allow sunspots and other solar surface features to be seen clearly.

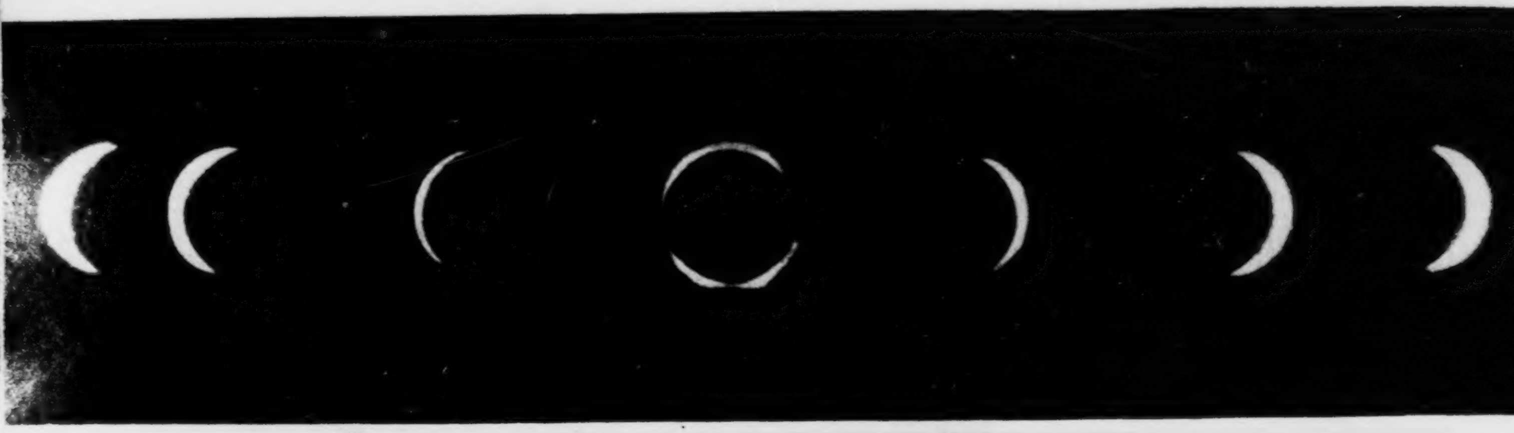
NCAR

Astronomers today are replacing the notion of a constant sun with one of an unpredictable, middle-aged star. Not only do sunspots blemish its surface, but there's evidence that the sunspot cycle does not have the dependable 11-year regularity ordinarily ascribed to it.

Recent observations have also disclosed unexpected solar oscillations ranging from five-minute wiggles to nearly three-hour pulsations. And satellites, on close inspection, have found the sun belching forth storms and eruptions far more frequently than had been suspected and developing holes in its luminous outer atmosphere. (See "The Sun, the Whole Sun," *Mosaic*, Volume 9, Number 1.)

Fading notions of the sun's constancy began to dim further three years ago when John A. Eddy, of the High Altitude Observatory in Boulder, Colorado, offered data showing that the sun was contracting, its diameter shrinking by about 0.1 percent each century. He set off a controversy that has yet to settle down.

The possibility of a shrinking sun arose as Eddy was searching historical records for possible sun-climate connections. He has been pursuing those connections for many years and is probably best known for his



1925 eclipse. Eclipse observers were stationed on rooftops on every city block along Riverside Drive in New York City to mark the edge of the 1925 eclipse shadow—a way to measure the sun's size.

Consolidated Edison Company of New York

rediscovery of the Maunder Minimum, a curious disappearance of sunspots that coincided with the cold extremes of Europe's little ice age in the 17th century.

In his work, Eddy happened across a set of historical data from London's Greenwich Observatory. Greenwich, the site of the world's prime meridian, had compiled centuries' worth of carefully recorded solar observations in marking local noon. To Eddy those data, especially the continuous series between 1750 and 1953, looked like "one of the longest continuous records in astronomy."

Eddy concentrated on observations made after 1850 with the Airy transit telescope installed at Greenwich that year. In those data he found evidence of an apparent two-arc-second diminution each year in the sun's horizontal diameter, a startling rate of shrinkage, equivalent to about a meter and half an hour along the sun's equator. The vertical diameter diminished less.

Eddy knew at the time that an 0.1 percent annual rate of shrinkage was preposterous over any extended length of time; in a hundred thousand years it would erase the sun altogether. And nothing in contemporary solar physics could account for such behavior. Eddy and his colleague Aaram Boornazian, a mathematician with S. Ross and Company in Boston, presented their findings to a meeting of the American Astronomical Society and awaited the fallout.

Response and retooling

Some researchers scoffed at Eddy's preposterous idea. Some went searching for evidence either to corroborate or refute the unsettling notion that the sun's size might be changing. Others pointed out errors and uncertainties in the information on which Eddy based his preliminary calculations.

According to current solar theory, a significant contraction in the size of the sun should cause a corresponding change in brightness or luminosity. What is uncertain, however, is how much the brightness would change, or even whether it would increase or diminish.

Eddy now believes his original finding was probably right in principle but wrong in the first estimate of the magnitude of solar shrinkage. He concedes that in interpreting the Greenwich records he underestimated the importance of two systematic effects: the condition of the atmosphere through which the sun was observed and the idiosyncrasies of the observers themselves. Neither factor is irrelevant.

Though the observing site was constant, Eddy acknowledges his mistake in assuming that London's air pollution did not vary or that changes in the brightness of the sky would not significantly alter the observations. In fact, the viewing conditions, a combination of pollutants that scatter light and atmospheric turbulence that makes the image of distant objects waver, did change. Greenwich looks out over an industrialized section of London where pollution got progressively worse until around 1890, then improved, then worsened again.

Pollutants change the sky's brightness, altering the apparent size of the sun against its backdrop. (This optical phenomenon, called irradiation, is a familiar one; it makes a person appear larger when wearing light clothing, or makes your feet seem huge when you don new white tennis shoes.) Against a clear, dark blue, unpolluted sky, the bright sun appears somewhat larger than when the sky's color, lightened by haze, diminishes the contrast.

On personal idiosyncrasies, John H. Parkinson, an astronomer with the University College, London, has dug into the historical records and found that the biases of some Greenwich observers "just leapt out" at him. "Although Greenwich was probably the best organized, best tooled-up observatory in the world at the time, and some of their men were brilliant at measuring, there were two... who simply couldn't get it right at all," he explains. Greenwich records carefully preserved each individual's personal idiosyncracies, some of which altered their perception of the sun by as much as 2.2 arc seconds.

Other problems

There are other problems in trying to wring a high degree of accuracy from the old Greenwich records. Among them is timing: A noontime transit only lasts about two minutes altogether. As Parkinson says, "The horizontal measurements were made by noting the length of time taken for the sun to drift across a wire in the eyepiece of the telescope. Before 1854, the observers watched the image of the sun while listening to the ticking of a pendulum clock. The precise time at which the sun crossed the meridian was then interpolated mentally between two successive ticks."

The time it took for the solar image to drift across the meridian-transit telescope's field was the measure used to calculate the horizonal diameter of the sun. A second's difference in recording the time could constitute an error of 15 arc seconds in the measurement. During the sun's transit, its vertical diameter was marked by adjusting setting circles and a micrometer attached to the eyepiece of the telescope. Positional astronomers used these measurements to help calculate the motions and orbit of the earth. Because the vertical measurements were taken with mechanisms subject to wear and change, Eddy felt they were probably less reliable than the horizonal ones.

Eddy knew from the outset that observers would have trouble determining exactly what constituted the edge of the sun as it passed the crosswires. Anyone who has seen photographs of the roiling motion on the sun's surface, the arcs and streamers boiling out of even a relatively quiet sun, can appreciate the amount of judgment required to declare that the edge of the sun's image has touched a certain point.

While the length of the Greenwich record would average out discrepancies, Eddy reasoned, he also went searching for analogous observations made elsewhere. He found them at the U.S. Naval Observatory in Washington, D.C. The observatory's records, also made of the sun at noon, did not extend as far back as Greenwich's. But Eddy was able to examine data gathered in the United States that overlapped chronologically those from England. He found that the Naval Observatory's transit information also indicated the sun's horizontal measurement had shrunk—by approximately the same amount as was found at Greenwich.

Parkinson expresses disdain for any meridian-transit measurements. He has visited the Greenwich telescope and tried to make the observation himself. "It's a horrific observation to make," he declares. "You have to squint at the right time through the crosswires and suddenly there's that awful thing jiggling and dancing about. A ragged, bouncing image. Measurements like that are so difficult to use; they're useless, really."

Indeed, if solar-transit observations were the only indicators available, finding historical changes in the solar diameter on the order of a few tenths of an arc second might be hopeless—much like trying to measure a millimeter with a yardstick. Yet the idea of significant changes was still intriguing and challenging, and others were encouraged to seek different astronomical archives: records of transits by the planet Mercury and measurements made during solar eclipses.

Mercurial transits

Mercury crosses the face of the sun about 13 times a century, a phenomenon observed by astronomers since the late 17th century. Mercury's transits appear to offer an advantage over noontime solar observations: There is more time to determine when the planet passes into the solar disc. But some confusion still exists about defining the edge of the sun; what astronomers call the black-drop phenomenon can make this observation tricky too.

As a transit occurs, one first sees a dark Mercury against a dark sky. As Mercury passes into the solar disc, it appears to trail a small drop of the sky's blackness with it. This optical phenomenon makes a clear definition of the edge of the sun difficult. It can cause different observers to declare different times when Mercury passes into the disc, even if they are standing side by side.

Astronomers write cautionary notes to forewarn observers about the black-drop phenomenon. Its duration can vary from ten seconds to several minutes after Mercury moves into the solar disc. Since transits are so infrequent, each observer, for all practical purposes, is seeing how the black drop behaves for the first time.

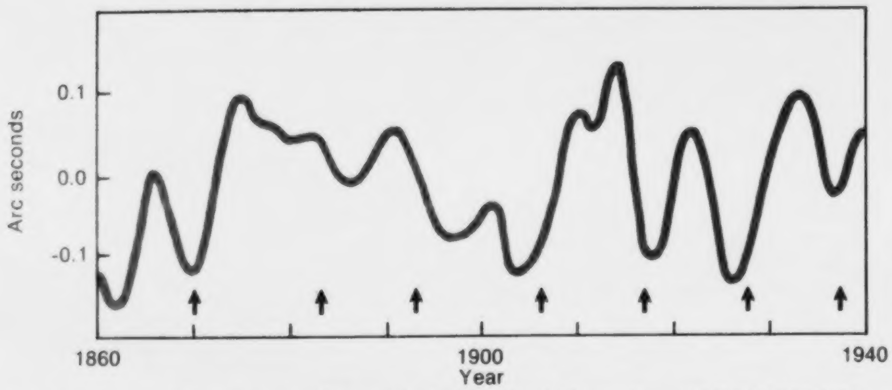
The more than five hours that tiny Mercury takes to skim across the solar disc does allow for more accurate timing than quick-look, high-noon observations. Nevertheless, of the more than 50 transits observed between the early 17th century and the present, there are only some two dozen that Irwin Shapiro, of the Massachusetts Institute of Technology, considers sufficiently accurate. Shapiro, a radar astronomer, is using very-long-baseline interferometry to explore the structure of quasars. As part of this work, he has collected historical records of Mercury transits, which he uses to determine the earth's motion. When the assertions of a shrinking solar diameter were made, Shapiro took his Mercury transit data and analyzed them as an independent measure of variations in the sun's size.

Shapiro's analysis showed that the sun was contracting, but by only 0.15 arc second in a century. Given the possible errors in the Mercury information, this tiny fraction, he argues, only shows there is no basis for assuming significant changes in the solar diameter. But Shapiro acknowledges that the sun's size probably oscillates with time.

Eddy, on the other hand, believes that a shrinkage of 0.10 an arc second over a century, the order of magnitude change he is now looking for, is significant. John Parkinson disagrees. He and his London colleagues, also drawing on records of Mercury's path, have calculated a shrinkage of 0.1[illegible] arc second per century. In addition, seven total solar eclipses he has studied seem to support his conclusions.

Shadow measurement

Did solar eclipse observations offer a better measure? Certainly more observers have corroborated them, since total solar eclipses are easily seen. The path of totality—the



Variation. By combining data from Greenwich, the U.S. Naval Observatory, and the Campidoglio Observatory in Italy, Ronald Gilliland plotted this variation of the solar diameter from 1860 to 1940. Arrows show the dates of sunspot maxima.

shadow cast on the earth when the moon moves between the earth and the sun—should yield a sensitive, though indirect, measure of the solar diameter. To make this observation accurately would require timed visual observations made just at the edge of the path of totality. Thus efforts to measure and time the shadow replace efforts to measure the pulsating sun itself.

Edmund Halley, of comet fame, organized elaborate measurements and timing of the 1715 solar eclipse as the path of totality passed over England. Sabatino Sofia of the National Aeronautics and Space Administration's Goddard Space Flight Center, looked into old records with colleagues David Dunham of the Computer Sciences Corporation and Alan Fiala of the U.S. Naval Observatory. They found that Halley enlisted the help of other enthusiasts to record observations that would bracket the shadow, producing data Halley hoped would confirm his predictions of the eclipse path. Some of those records, with appropriate caveats, can be important to solar-dimensions research today.

Why the caveats? Crucial to the accuracy of Halley's information, for example, is knowing exactly where Theophilus Shelton, Esq., stood in Darrington, Yorkshire, to observe the northern edge of the eclipse shadow on May 3, 1715. A search of old records located Shelton's position, to Sofia's satisfaction, to within 200 meters. But to be satisfied, one must accept Halley's word that Shelton was actually in Darrington proper and not some distance outside the town.

Another intriguing record is that for the eclipse of January 24, 1925, when the path of totality was over New York City. At that time the editors of *Scientific American* encouraged people all over the eastern United States to gather information. One response came from an unsung hero employed by the electric companies affiliated with the Consolidated Gas Company of New York (now Con Edison). He set up a network of eclipse observers stationed on rooftops on every city block along Riverside Drive. This yielded a measure of one edge of the eclipse path accurate down to a city block. Whoever organized this public-spirited effort, however, subsumed himself in the corporate image; his identity is lost somewhere in the files of Con Edison's corporate grandfather.

Lost data

Sofia and his colleagues have tried to reclaim all the data gathered on the 1925 eclipse. Ernest W. Brown of Yale Observatory might have been their best hope. Brown received all the information on the 1925 eclipse path from volunteers who filled in the observers' coupons printed in *Scientific American* and in daily newspapers. He also had a plethora of notes and letters from people fascinated by the eclipse and its rarity.

Brown published some of his findings in the *Astronomical Journal* in 1926. In his article he noted that a good deal of information had been reported about the edge of the eclipse shadow, but he had not yet analyzed it. He apparently never got to it before he died.

Sofia asked about the shadow observation and was told that the records hadn't been seen. When the Yale Observatory was physically moved, cartons of Brown's data apparently were discarded. In the hope of being able to salvage something, Sofia took a trip to New Haven "to go through all the boxes and every piece of Brown's mail" that could be found. He turned up nothing.

Sofia and his colleagues have combined historical information with data gathered from two recent eclipses: the 1976 event over Australia and the 1979 eclipse across North America. While they plan more eclipse observations in the future, they now calculate that between 1715 and 1979 the solar radius decreased by about 0.34 arc second. Irwin Shapiro, however, thinks that Sofia is "skating on extraordinarily thin ice," especially with regard to the 1715 eclipse information. Shapiro calls the delving into old astronomical records "a veritable new cottage industry among solar physicists, each producing results more embarrassing than the last."

John Eddy agrees that the historical information "has been pretty well pushed to its limits." But, he adds, "I think we've cracked an old prejudice about the size of the sun. And if the sun's diameter really does change, so should its brightness, which could well have important climatic effects."

Oscillations

Ronald Gilliland, a researcher with the High Altitude Observatory, declares that "although there are questions about every one of these data sets, and quibbles about the exact amount of shrinkage each shows, the important thing is they all show a decrease in the diameter of the sun." While each kind of data shows differing rates for that decrease, Gilliland believes the inconsistencies stem from the fact that there is not a steady decline in the solar diameter but subtle oscillations in size. The difference between the Mercury transit results and the eclipse results could be explained in relation to a larger cycle of solar oscillations, he says. And he adds that apparent inconsistencies in the amount of observed shrinkage can be explained by the fact that the diameter of the sun is oscillating on a cycle of approximately 76 to 80 years.

To reach this conclusion, Gilliland examined data sets of the sun's history covering 258 years. His analysis shows that most Mercury transit observations were made during the sun's maximum phase, while more accurate contemporary observations are being made on the negative, or shrinking, side of the phase.

London's John Parkinson, on the other hand, is aware of a possible periodicity but considers it of no consequence. He cites two different papers he is currently reviewing. Each shows the same wave and the same amplitude for 80-year cycles, he says, "and each is exactly 180 degrees out of phase with the other."

Parkinson stresses the need to be skeptical, to be critical of any observations you didn't make yourself, and the importance of "getting inside the guy's head, figuring out why he's making it and what tools he had." He cites the example of the writings of a monk who timed the solar eclipse of 1415 to two repeti-

tions of the *Miseri Mei,* the 51st Psalm. Today, if a researcher were to come across that bit of information, it would be a common sort of error to time how long it takes to repeat the 51st Psalm twice—in English. And it doesn't work. "But if you recite it in Latin, you get about 220 seconds duration for the eclipse," Parkinson says, a number consistent with a constant solar diameter.

In his view, all reported discrepancies result from the observational method, not variations in the sun. "The diameter you get with meridian transits is not the same as with eclipse measurements. When you use meridian transits, the sun is 1.5 arc seconds too big."

Taking this into account, observers are currently establishing several independent programs to examine the sun for small-scale changes. The High Altitude Observatory has installed a modern meridian-transit telescope to time the sun's noon crossings of the sky electronically. Using several banks of hundreds of independent photoelectric detectors to measure both the vertical and horizontal diameters of the sun will, according to the observatory's Timothy Brown, "give a picture of a slice across the sun 30 times a second." Each array of more than 500 diodes will record the solar diameter by measuring the intensity of individual points of light and the time the intensity is recorded. Each noontime transit will provide thousands of bits of data.

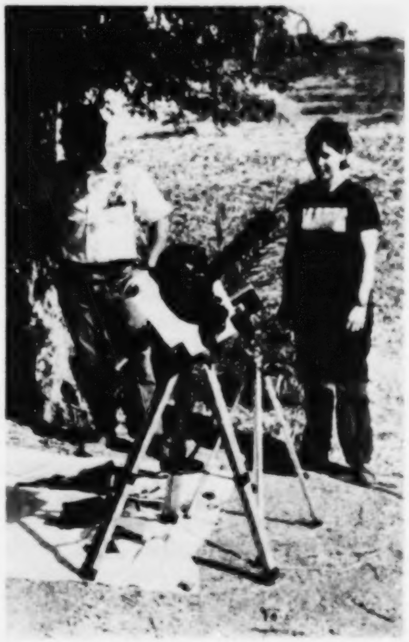
Defining the edge

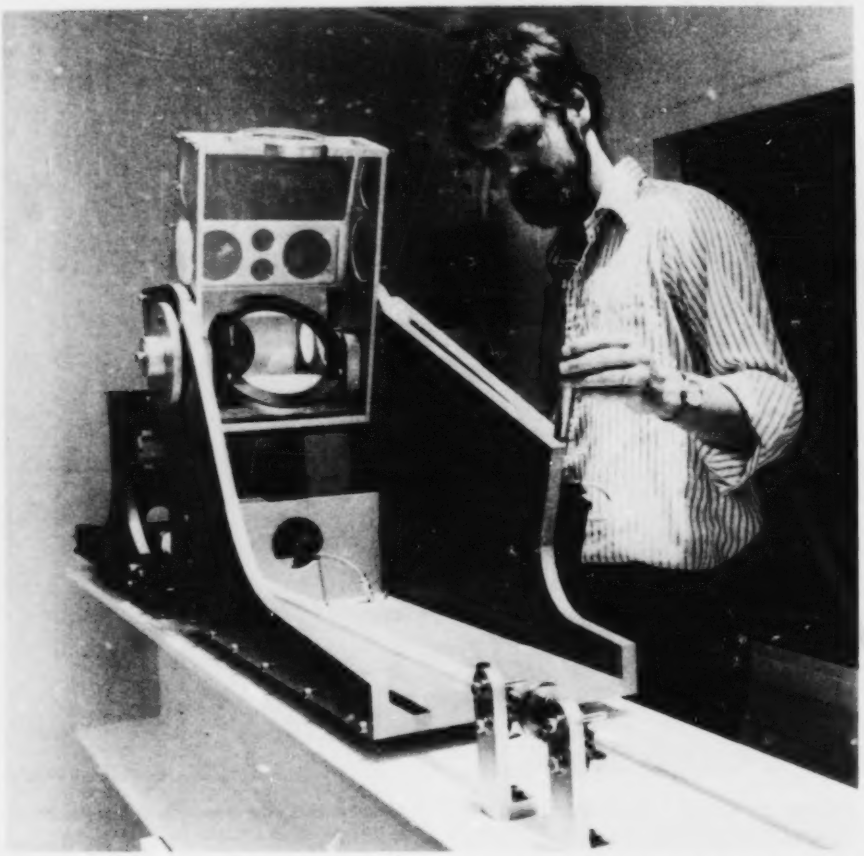
According to Brown, the difficult problem of finding the edge of the sun was solved by Henry Hill at the University of Arizona, who used a mathematical technique to calculate the brightness distribution. Hill's technique has "nice properties," Brown says. "It doesn't care about how bright the sun is, or about background light, or particle scattering and atmospheric seeing."

Hill, best known for his work on long oscillations in the whole sun, will be extending his work on solar variations to see what light his measurements can shed on the question of changes in the solar diameter. He wants to find out whether long-term changes in size involve the whole sun, or, as most astronomers today believe, only the convective zone and the photosphere.

The Kitt Peak National Observatory's huge vacuum telescope in Arizona has also been pressed into service. Since December 1979, researchers there have been making observations in near infrared wavelengths as the sun drifts across the telescope's field of view. According to Kitt Peak solar astronomer Harrison Jones, "It will take about

Observation. A research team monitored the solar image projected onto a cardboard screen during the solar eclipse of February 16, 1980, in India. Alan Fiala (seated, center) was assisted by Joan Lindell (left) and John Sparrow. Maurice Quidley and Joan Dunham (below) assisted.





Noon transit. Timothy Brown adjusts the modern meridian-transit telescope at the High Altitude Observatory. The device times the sun's noon crossings of the sky.

NCAR

five years of data gathered once a week before we can say anything about whether the sun has decreased."

"Problems with new methods require 10 to 15 years of regular observations," cautions astronomer Alan Fiala, an astronomer at the U.S. Naval Observatory, since it may take that long to determine what is actually being measured. To get around this problem, John Parkinson only half-jokingly suggests using the Airy telescope that made the Greenwich observations. "It's still around. Let's get the creaky, cranky old instrument out and try to do it again."

Better questions

Writer Norman Cousins once observed that science at its best provides us with better questions, not absolute answers. This definition seems to describe John Eddy's question about the solar diameter. Not only must one now ask whether the sun's size is constant, but whether any changes can be measured, how they can best be detected, and how reliable the astronomical records of history are. Even fundamental ideas about how the sun works may have to be questioned, which in turn could raise new queries about stellar evolution.

And shimmering in the distance is yet another question which, Eddy acknowledges, has "an almost hypnotic attraction." That is the tantalizing possibility that even small changes in the solar diameter might be felt on earth as subtle changes in solar luminosity, and hence might influence climate.

Questions about the evolution of the sun and other stars in our galaxy are fascinating. But the climate connection has a particular urgency. As Samuel P. Langley, a pioneer in American solar astronomy, once observed: "Every nebula might be wiped out of the sky tonight without affecting the price of a laborer's dinner, while a small change in the solar radiation may conceivably cause the deaths of numberless men in an Indian famine." The urgency of the search has only increased, not diminished, with time. •

The National Science Foundation contributes to the support of the work discussed in this article through its Solar Terrestrial Program.

The Gender Changers

by Ben Patrusky

Many lower-order plants and animals are hermaphroditic. The evolutionary and environmental conditions that trigger sexual elasticity are becoming clear to theoreticians.

Hermaphroditism is defined as the presence in the same plant or animal of both male and female reproductive organs. From an anthropocentric point of view, it smacks of the unnatural and the bizarre. But many lower-order plants and animals, including some vertebrates, manifest a sometime sexuality. This can be synchronous hermaphroditism, the ability to produce egg and sperm, seed and pollen, at the same time. Or it can be sequential hermaphroditism, a switchover from one kind of gamete production to another.

Some of the species of shrimp on which humans feed, for example, start life as nothing but reproductive males and shift to a predominance of females as populations mature. Some marine vertebrates, including several species of fish that inhabit the waters around coral reefs, do similar sexual flip-flops.

Trickier still are the sex changes undergone by flowering plants. Most are synchronous hermaphrodites, fringed with flowers that possess both stamens and pistils, the pollen-making male and seed-yielding female organs. Some, however, are one sex one year, another in an ensuing year, then back to square one in yet a third flowering season.

What gives rise to these flexible mating dispositions? What advantage, if any, does such sexual inconstancy confer? Scientists finally seem to be coming to terms with this long-standing evolutionary mystery. Over the last decade, as a result of a sudden step-up in research efforts to understand sexual elasticity, there is now emerging a clear picture of the reasons for and the circumstances that favor these changes.

This increase of interest in hermaphroditism stems directly from an idea first introduced more than a decade ago by biologist Michael Ghiselin, now at the University of Utah. Ghiselin's proposition drew on a fundamental precept of evolutionary theory: that natural selection acts on individual organisms, not on their species. That notion is fundamental to traditional Darwinian theory, yet until Ghiselin, it seems that the only way evolutionists could make sense of hermaphroditism was to consider the evolutionary advantage it might offer to a species.

One gender at a time

To arrive at what has come to be called his size-advantage hypothesis, Ghiselin reasoned that if fertility varies from sex to sex as size changes, an individual should be whatever sex has the highest fertility at any given size. Thus, for example, if large size increases the egg output of a female but does not give large males any advantage in the competition to fertilize eggs, then it would be advantageous for an individual to operate as a male when small and as a female when older and larger, as long as the option is available. If large size gave males a reproductive advantage, the reverse order of sex change, female to male as size increases, would be favored. Ghiselin thus offered sequential hermaphroditism as one of nature's ways to maximize an individual's fecundity over its lifespan.

Publication of the size-advantage hypothesis stirred the interest and imaginations of several research biologists, among them Robert R. Warner, now at the University of California at Santa Barbara, and Eric L. Charnov of the University of Utah. Using the principles of population genetics, each went on to enlarge upon and formalize the Ghiselin precept, producing comprehensive mathematical models that qualitatively defined the circumstances favoring sex reversal, the direction of change, and the optimum switching age.

Because their research goals were somewhat different, the two took somewhat different tacks in testing their propositions. For Charnov, hermaphroditism was a special case. His overriding interest was the general question of sex allocation; he was seeking to learn how natural selection operates to manipulate sex ratios among progeny so as to maintain optimum balance in a species between sperm and egg production. Warner, a practicing marine biologist concerned primarily with animal behavior, applied his mathematical construct to exploration of the social situations that gave individuals a size-related mating advantage.

A flexible tau

Charnov elected to test his formal version of the size-advantage hypothesis by, as he put it, "solving for tau," his designation for the age at which sex reversal occurs. What he did in effect was predict the results of several natural experiments that, according to his model, would produce significant changes in tau. Charnov believed the age of sex change in hermaphroditic species was not fixed; tau, he said, was elastic, a product of pressures that could affect a population.

The populations around which Charnov built his model were the members of the shrimp family Pandalidae. Made up of about 20 species that inhabit temperate-to-subarctic waters in the Northern Hemisphere, pandalids are the prawns harvested for human consumption. They begin life as breeding males and reverse course later. Technically, these shrimp are designated sequential protandrous hermaphrodites. Protandry is maleness at the outset; its reverse is protogyny.

One prediction devolving from Charnov's construct concerned the way heavier fishing of the shrimp beds would affect tau, the switching age. He figured that as more and more shrimp were taken, tau would shift to a smaller and smaller value—a younger and younger age. As Charnov explains it: "Breeding takes place in the fall." After copulation, females carry the eggs until they hatch in the spring. The larvae are free-floating plankters at first, settling to the bottom in mid- or late summer as they mature. The shrimp breed for the first time a year and a half after hatching and at yearly intervals after that. A population of shrimp with a lifespan of five-and-a-half years would comprise four breeding groups.

"Left undisturbed, members of such a shrimp population would normally spend the first year reproducing as males and the last three as females. But the sudden impact of a harvest is that it produces the effect of a sharp increase in mortality, a reduced lifespan. Shorter-lived shrimp would no longer grow to be as large as they otherwise might have, and there would be fewer breeding cycles. Under these circumstances, instead of spending one year as a male and three as female, there should now be a downward shift and we should see smaller females in the population."

Soon after making this prediction, Charnov gained access to data collected from Danish and Swedish fisheries over the last four decades. The data showed that between 1930 and 1956, the northern European shrimp catch was relatively stable from year to year. From 1956 to 1960, however, the catch was increased from 2,070 metric tons to 5,670 metric tons. During the earlier, stable period, no females had been found to be less than 75 millimeters in length. But when the haul was abruptly increased, the proportion of females under 75 millimeters increased significantly.

The implication was clear to Charnov: Tau had been reduced. The shrimp catch imposed the increased mortality, and that changed the optimum size at which a sex change would confer reproductive advantage. The response was for the shrimp to begin breeding as females at a younger age.

Testing the hypothesis

Other explanations, meant to discount the notion that fishing had a role in changing tau, have been put forth. These include a suggestion that smaller female shrimp may have migrated from elsewhere or that a reduction in the food supply available to the shrimp produced a slower rate of growth; thus females, growing more slowly, would appear to be younger than they were. Charnov finds this unconvincing, and confirmation of another prediction derived from his mathematical model seems to bear him out.

He argues that shrimp from different latitudes should demonstrate different taus, different switchover ages. Accordingly, Charnov forecasts, shrimp from more southerly latitudes would spend less time as males (tau would be smaller) than would those from more northerly locales, and southern female prawns would be smaller than northern females. His reasoning: Shrimp living in colder waters have longer lifespans than those in warmer climes; lifespan increases from three years to eight years as latitude increases; as lifespan goes, so goes tau.

Charnov put his prediction to the test by examining data from 28 different shrimp populations living in locations from the Gulf of Maine to the Arctic Circle in the Atlantic and from the coastal waters of California to the Bering Strait in the Pacific. His findings supported his suppositions; tau tracked with latitude.

Charnov proposed one other test. He suggested that tau would also change in relation to the breeding success of a given shrimp population. As he explains, survival of shrimp larvae may vary from season to season. During years of high breeding success, many more young, small shrimp will join the population, driving down the average age of the community; the reverse would be true in a string of bad years. Charnov's mathematics argues that the younger the average age—the greater the ratio of younger, smaller breeders to older, larger ones—the lower tau would be. Hence, the fraction of smaller shrimp reproducing as females would be greater, too. The opposite would be true for poor breeding years; tau would increase. Analysis of recent data from California and Oregon fisheries suggests just such year-to-year fluctuations in tau, Charnov reports.

Social factors

Charnov's shrimp are protandrous hermaphrodites. There are also protogynous hermaphrodites, those that shift from female first to male later, and the direction of change has much to do with social constraints. The thrust of Warner's research has been to apply his model to the study of social factors that influence sex reversal.

It is generally accepted, for example, that a social system featuring random mating—any individual with any other individual, regardless of size—would favor protandry. But where large females pair with large males and small females with small males, no sex change would be likely; individuals would be dioecious (one sex for a lifetime). And in systems in which large males monopolize mating, the circumstances are ripe for protogyny.

The cleaner wrasse, a fish found in Australia's Great Barrier Reef, is a case in point. Named for the valet service it performs—cleaning the skin of other fish—this species affords its males up to eight times the fertility advantage of females: one large male mates with six to eight females daily.

Theory would predict protogynous patterns for the cleaner wrasse—females when small, males when large. And observation shows this to be the case. Indeed, studies by D. Ross Robertson of the Smithsonian Tropical Research Institute, based in Panama, were the first demonstration of the way social forces governed sex reversal. Robertson noted that when the dominant male was removed, the largest female in the harem changed sex almost immediately and took on the departed male's role.

Robert Warner has gone into the field to put theory to test. The primary object of his study was another species of wrasse, the bluehead, found among the coral reefs of the San Blas Islands off Panama's Caribbean coast. Over a period of several years, Warner—in collaboration with Robertson and Egbert G. Leigh, Jr., another Smithsonian marine biologist—made numerous snorkeling expeditions to watch and photograph his research favorites in their native habitat.

Warner has described two distinctively different mating systems within the same breeding colony, with population distribution dependent on reef size. The smallest reefs, for example, were composed almost exclusively of protogynous hermaphrodites arranged in a three-tiered social system: large, terminal, territorial bluehead male spawners, identifiable by their coloring; nonaggressive, nonterritorial, floater males; and reproductive females. As reef size increased, these hermaphrodites were found interspersed with primary males, those that remained male throughout their lifespans. The larger the reef, the larger the complement of primaries.

Warner says his mathematical construct helped him make sense of the environmental and social circumstances that gave rise to this curious mixed sexual bag. The phenomenon is explicable in terms of what he calls "the penalty of excess selfishness," and he is able to predict, merely on the basis of reef size, the distribution of the various populations.

Small reef—big reef

Each of the hundreds of San Blas reefs, ranging from desk- to building-size, maintains its own closed colony of bluehead wrasse. Forming their own interacting social unit, adult members of each reef have no significant contact with colonies of other reefs.

On a typical reef, spawning occurs for a few hours each midday throughout the year. At this time, most of the sexually active females gather at specific sites on the reef, generally along the outer rim on the down-current side. There the large, terminal males set up temporary spawning territories. These sites are probably chosen, Warner explains, because there is a greater chance that fertilized eggs released there as planktonic larvae will be swept off the reef and away from egg-devouring coral.

Females ready to spawn generally choose the largest and most brightly colored males. The females will often delay mating if the largest males are otherwise engaged, rather than mate with a smaller territorial male.

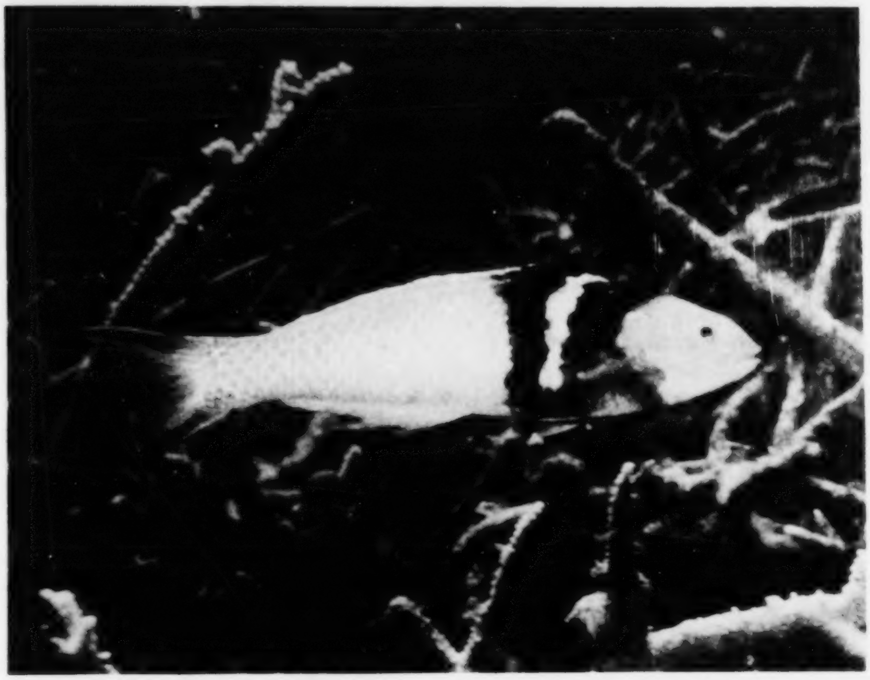
Fertilization is external; egg and sperm are released at the apex of an upward, side-by-side mating rush lasting no more than two or three seconds. On the average, at least on small reefs, large males repeat this mating behavior about 40 times a day. By contrast, females spawn but once a day, spending most of their time feeding elsewhere along the reef.

Under these circumstances, and as the theory derived from size advantage clearly predicts, protogynous, sequential hermaphroditism is the order of the day. "It's worth it for some larger females to take a chance on changing sex," says Warner, "because just one day as a spawning male can make up for a month or more of days lost as a reproductive female." The disproportionate reproductive success available to the larger males also explains the presence of floaters, which Warner describes as "young bucks waiting in the wings," ready to take over a spawning territory should a dominant male vanish from the scene.

As with Charnov's pandalid shrimps, the time of sex changeover, tau, is not fixed for blueheads by arrival at a preset size; it depends instead on the social conditions that prevail locally. To demonstrate, Warner removed larger, territorial males from the reef. That caused the next largest male to shift from floater to territorial spawner. It also induced the largest females to reverse sex. By taking more and more of the larger members away in laboratory populations, Warner was able to drive tau down to a smaller and smaller value—an earlier and earlier age of conversion.

"The fish respond to environmental opportunity," notes Warner. "But in all cases males mate with smaller females." Once the protogynous switch takes place, females-become-males generally join the pool of non-spawning floaters, although they could, if circumstances allowed, move in right away to take over a territory. No previous experience as a male is necessary.

Very little is known about the internal hormonal changes that bring about the re-



Hermaphrodites. The bluehead wrasse (above) and shrimp of the Pandalidae family are hermaphrodites studied to determine the conditions that trigger sexual change.

Robert Warner, D. R. Robertson, and E. G. Leigh/*Science*, 14 November 1975; Eric Charnov

structuring of reproductive plumbing required to convert wrasses from egg makers to sperm makers. But there is some recent evidence from other fish species that offers a clue to one possible trigger. It has to do with the amount of aggression the female experiences from the male. Larger females, it seems, are subjected to more aggressive male behavior; a decrease in male aggression, as would be the case if the male vanished or died, spurs the sex change.

Complex arrangements

Bluehead sexual dynamics grow more complex with increasing reef size. With larger reefs come larger wrasse populations—up to a thousand times the size of small-reef groups. But mating-site availability does not keep pace with the growth in population. This results in a relative scarcity of large, terminal-phase males in large populations and increased mating opportunities for those that are present.

Surveys of the reef populations bear this out: Territorial males make up 10 to 25 percent of the population when the reef is small but less than 1 percent when it is very large. On some reefs, dominant males may spawn more than a hundred times a day. Ultimately, however, large-reef territorial males begin to pay what Warner calls the penalty of excess selfishness. With the proportion of females to terminal males increasing with larger reef size, says Warner, "there is opportunity for, and even a premium on, finding another way of being male—one that will reduce the penalty of youth and small size." What has evolved, as a result, are alternative, or non-sex-changing, cheater males that take advantage of the overabundance of females.

These small, primary males circumvent the standard mating ritual. While the female waits her turn to spawn with a territorial male of her choice, she is susceptible to sneak gang attacks by small, young males. These males will band together, intercept a ripe female on the way to a mating territory, pursue her until she relinquishes her eggs, and then spawn as a group. Further, one of the small males will sometimes intrude on a territorial male-female pair in its mating rush. The intruder will release its sperm into the water at the same time the mating climax occurs, injecting its sperm into the fertile cloud.

With increasing reef size, management of females by territory-protecting dominant males becomes progressively more difficult. Young primary males ("little more than sperm-making machines," in Warner's words) succeed more and more with their interference and group-spawning tactics. By contrast, on small reefs terminal-phase males should be better able to patrol their spawning domains, by preventing these alternative males from taking root in the colony. Indeed, primary males were found to make up less than 2 percent of individuals on smaller reefs, compared to 20 percent on the large ones.

On the biggest reefs, small-male population density at the mating sites was so great that group spawning predominated. At least 80 percent of the females were snatched up by the primary males, while the large, territorial males were relegated to less-than-the-best mating sites. The relationship of reef size to mating style was so clear-cut that Warner was able to predict the proportion of sex changers to nonchangers in the bluehead wrasse population merely by comparing reef dimensions gleaned from aerial photographs.

The wrasse studies, says Warner, lend strong support to the general theory that where one sex gains tremendous fecundity advantage over another because of size, sex change is adaptive—i.e., natural selection will likely favor genes that promote sequential hermaphroditism. The corollary: where size confers no such advantage (smaller spawning males do as well as larger ones) hermaphroditism confers no selective advantage and sex change will not occur.

Other species

Tested against other marine species, the size-advantage model has been found to work well in either of two ways. It predicts the nature of the social system once the sexual proclivities of the species are known. Conversely, it permits prediction of the presence of sex changing in a species on the basis of a few facts known about the prevailing social system.

There is a species of wrasse found off the coast of England, for example, in which the male makes a nest and spends a great deal of time caring for the clutch. Warner's theory-based assumption would be that this species is probably dioecious—made up of non-sex reversers. The reason is that nesting reduces reproduction time and the maximum number of matings possible; the male is also too busy to keep other, smaller males from spawning. Under these circumstances, being a large male doesn't provide any fecundity advantage.

Then why not protandry? Warner speculates that in this species large females would elect to mate with large males, since they should be best equipped to protect and tend to the nest. That would leave small males to mate with small females. And as theory has it, where there is no size advantage accruing to either sex, then there is no evolutionary incentive to hermaphroditism. And what do the field studies of this nest-building wrasse actually reveal? A species of lifelong males and stay-put females.

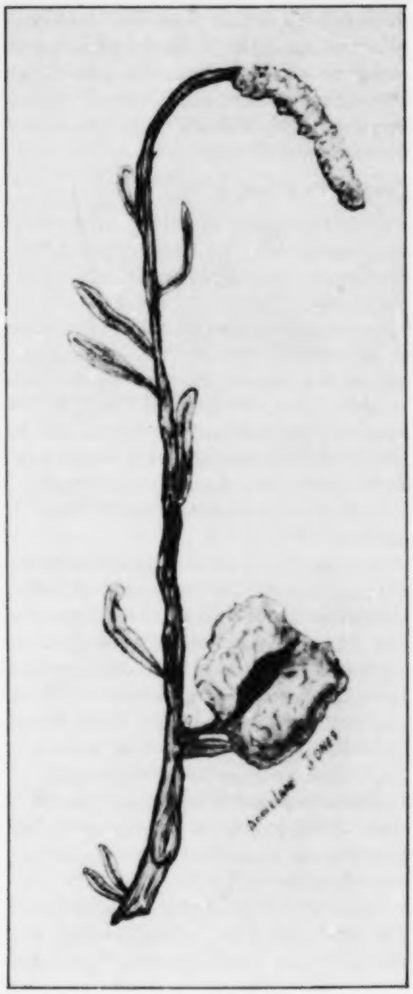
Another example is the clownfish, which takes up residence in sea anemones. Here the mating group is limited to two unequally sized adults living monogamously and mating, only with each other, once a day. When one of the adults disappears, one of the juveniles takes over as mate. Given these circumstances, theory would predict a sex-changing system of the protandrous variety: male when small, female when big. And that in fact is what actually exists. The larger of the two in the mating pair is invariably female, which maximizes the pair's fecundity. When the female is removed, the male of the pair changes sex and a juvenile slips into the slot left vacant by the transformed male.

There are many higher organisms that seem to fit the criteria for sex change offered by the size-advantage hypothesis. Among deer, sheep, and goats, for example, older males monopolize a disproportionate share of matings. But there is no hermaphroditism in higher vertebrates. Indeed, among all vertebrates only certain species of fish are known to practice sex change.

Warner suggests several reasons for this. For one, the cost of changing the reproductive plumbing of terrestrial animals, which engage in internal fertilization, may far outweigh the benefits to be derived from hermaphroditism. Moreover, animals may need not just size but experience to fight successfully for females under conditions of intense male competition. A large animal that had just turned male, says Warner, would be at a hopeless disadvantage relative to his more experienced fellows, despite his size.

Plants

Not all terrestrial organisms have forsaken hermaphroditism, however. Certainly when it comes to plants, land-based is hardly a synonym for sexual stability. In fact, if sexual option serves to enhance fecundity, as measured in terms of offspring produced in an individual's lifetime, then the world's flora are truly in reproductive clover: Of all the extant species of flowering plants, some 93 percent live a sexual double life as synchronous, or simultaneous, hermaphrodites. In most species, both male and female sex organs—pollen-producing stamens and ovule- or seed-bearing pistils—exist side-by-side in the very same flower. In the remaining 7 percent, all flowers house either stamens or pistils alone. And therein lies an extraordinary tale of hitherto unsuspected turnabouts in



Plants. Typical morphology of a hermaphrodite *Atriplex canescens* is shown above. The upper portion represents male flowers; the fruit is from a female flower. *A. canescens* in Kingston Canyon, Utah (right) are male biased. Researchers have detected a striking association between environmental fluctuations and sexual expression.

D. Carl Freeman and Durant McArthur

botanical sexuality.

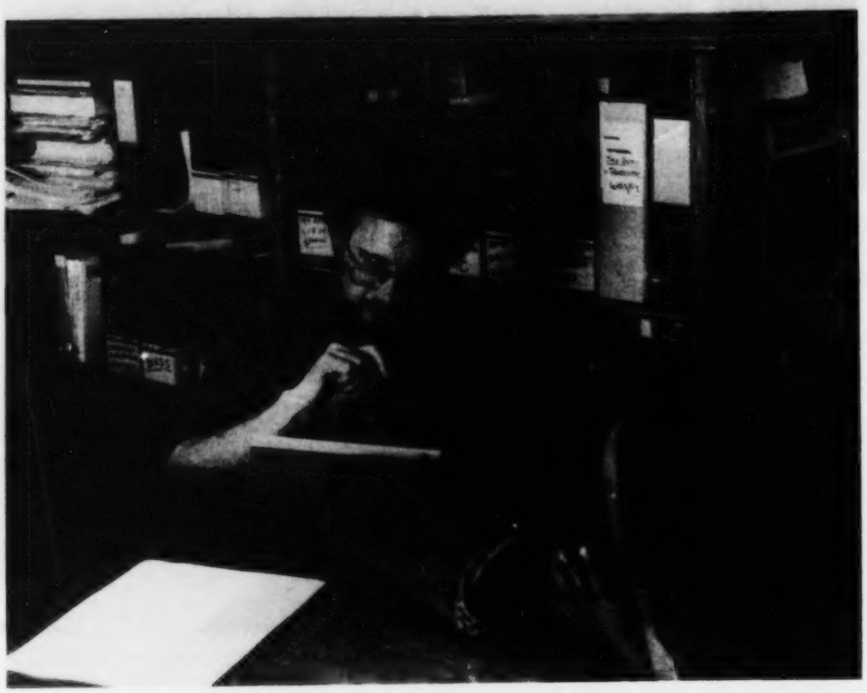
Plants with single-gender flowers are generally designated as either dioecious or monoecious. In dioecious species such as spinach, hemp, aspen, and box elder, all flowers on any individual plant are of the same sex. Monoecious species, such as corn, cucumber, and squash, produce both kinds of single-gender flowers—staminate males and pistillate females—on the same plant. Among the world's terrestrial flora, monoecious species are about twice as abundant as the dioecious kinds.

It has long been considered likely that dioecious species had evolved from simultaneous hermaphroditic ancestors. But no one could say why. Darwin, who did important work on the ecology of dioecious plants, was unable to figure out what advantage devolved from unisexuality; nor could a long line of plant biologists who followed.

The problem really came down to numbers. Scientists reckoned that in giving up simultaneous hermaphroditism and going dioecious, plants suffered at least a twofold penalty in reproductive success. Put another way, the same number of flowers on a simultaneous hermaphrodite yielded twice as many offspring. If plants were willing to pay the price, they must be receiving some benefit. But what? How was unisexuality adaptive to plants?

Answers

There was no answer until 1940 when Karl Mather, a British botanist, published a paper that seemed to provide one. Mather argued that the dioecious state had evolved as a mechanism for the promotion of outbreeding. He recognized that many bisexual flowers could pollinate themselves and that



Eric Charnov. The University of Utah research biologist tested hermaphroditism in shrimp by "solving for tau," Charnov's designation for the age at which sex reversal occurs.

University of Utah

such inbreeding could have disastrous effects on the fitness of subsequent generations. By becoming dioecious, Mather contended, plants improved the genetic quality of their offspring enough to compensate for the twofold loss in quantity. Implicit in this explanation was the assumption that a genetic arrangement existed that allowed for the expression of one and only one sex by a dioecious individual and hence made outcrossing mandatory. (See "Preventing incest at pistil point," *Mosaic*, Volume 12, Number 4.)

Mather's proposal took root, becoming the reigning theory and botany's party line for more than three decades. But in the mid-1970s, Kimball T. Harper of Brigham Young University and D. Carl Freeman, now a plant ecologist at Wayne State University, helped turn up new experimental evidence that posed a serious challenge to the Mather doctrine. Not that the outcrossing theory didn't have merit, but in their view that was far from the full explanation for the evolution of dioecy. They took direct issue with the idea that dioecy meant a fixed-forever gender.

Environmental influences

Freeman and Harper contended that individuals of many dioecious plant species were actually quite flexible sexually and had the wherewithal to change sex in response to changing environments. As they saw it, dioecy was really only a one-sided reflection of a hermaphrodite exercising an option. The sex it manifested presumably gave it a fecundity advantage over other plants growing in the same environment. By the same token, if the environment changed in ways that gave seed-producing plants an advantage, then dioecious species in a masculine phase could flip gender and produce pistil-bearing flowers.

Their argument did not immediately take hold; most of the college botany texts introduced through the 1970s made no mention of any possible role of environment in determining the sex of dioecious plants. And yet, long before Freeman and Harper, ample reason to suspect otherwise had already appeared.

In the 1920s, for example, John H. Shaffner, a botanist at Ohio State University, did a landmark study on the life cycle of the jack-in-the-pulpit, *Arisaema triphyllum*, a short-lived perennial found in the eastern United States. This plant, he discovered, was a sequential hermaphrodite. During its first year, *A. triphyllum* flowered as a male; with ensuing years, a few female flowers began to appear. As the plant aged, the population of female flowers continued to increase until, ultimately, sex reversal was complete.

In later trials, Shaffner manipulated the pseudocorm, an underground structure that serves as the jack-in-the-pulpit's metabolic reservoir. He found that he could alter sex from one flowering season to the next. If he started with a female, for instance, and cut off a fraction of the corm, the plant would flower as a male the following year. If he allowed the corm to increase in size again, the plant would ultimately revert to its former dioecious and all-female state.

Lysenko's shadow

In another series of studies involving a dioecious species called *Thalictrum dioecium*, Shaffner deprived all-female plants of normal allotments of water and fertilizer. In the next season, those same plants flowered either as all-male or as monoecious. In a series of now-classic papers, Shaffner argued persuasively that environment could alter a plant's sexual proclivity. Unfortunately, he published at a time when the debate over Lysenkoism was coming to a head; Shaffner was caught in a maelstrom not of his own making.

The Lysenko doctrine, formulated by the Soviet agronomist Trofim Lysenko, proposed that genes could be modified by experience and that acquired characteristics could be inherited. The position was politicized and Lysenko prevailed in post-World War II, Stalinist, Russia, sparking a fierce debate into the 1950s. Lysenko was ultimately discredited. Shaffner's observations, superficially reminiscent of the dispute, were given short shrift, passed off as little more than experimental artifacts and of no real importance in nature.

In the 1960s, plant physiologists found that they could modify the sex of dioecious species by administering newly discovered plant hormones in varying proportions. But bowing to the prevailing wisdom of the geneticists and the dominant canon of immutable sexuality, they failed to link their results to what was happening in the real world.

From time to time, other examples of apparent botanical sexual inconstancy surfaced. But geneticists were quick to explain away what they called intersexes, or chromosomal mosaics, as inconsequential rarities in nature. Given this doctrinaire climate, it is small wonder that not much note was taken of a 1966 report by Fred Vasek, a taxonomist at the University of California at Riverside. As part of other studies, Vasek had marked two different species of dioecious juniper with numbered tags and recorded their sex. Over a period of two to five years, more than 20 percent had changed sex. What was especially significant was that this represented evidence of sex reversal in the wild.

In 1973, Katherine B. Gregg, a doctoral

candidate at the University of Miami, altered the sex of tropical orchids by varying light intensity. In diminished light, the equivalent of shade, the orchids bloomed as male. Placed in bright light, males became monoecious or female. Similarly, females that were moved into shade drifted toward maleness. Gregg's laboratory results duplicated those of studies done a decade earlier on orchids in their native tropical habitat by C. H. Dodson, now of the Selby Botanical Gardens in Sarasota, Florida. Gregg confirmed that in shaded areas, the orchid population was predominantly male, but where light penetrated through gaps in the thick forest, the plants yielded female flowers.

Inescapable conclusion

Well versed in the literature of plant genetics and plant physiology, D. Carl Freeman, while still a graduate student at Brigham Young University, felt he could draw only one conclusion from this train of extant but ignored evidence: Some individuals of dioecious plant species could alter their sex in response to environmental change. He made his conviction known to Kimball Harper, who in turn contacted Durant McArthur. McArthur was a forest-service geneticist at the Shrub Sciences Laboratory, run by the United States Department of Agriculture in Ephraim, Utah.

Since 1972, McArthur had been observing the year-to-year development of 672 desert shrubs of the dioecious species *Atriplex canescens*. He was working up his data on sexual lability when he and Freeman first met. His data showed that the plants' expression of gender had varied from season to season; virtually every variety of sex change was manifested.

Subsequently, McArthur, joined by Freeman, detected a striking association between environmental fluctuations and ensuing sexual expression. Following unusual environmental stress, including an especially severe winter and a prolonged period of drought, there was a wholesale shift toward maleness. When stressful conditions lifted, the proportion of females in the population increased. In all, more than 40 percent of the plants changed their sexual state during the several years of observation. McArthur and Freeman have since concluded that *A. canescens* has a sexual system that produces constant males, constant females, and labile plants that may be male, female, or hermaphroditic in different years in response to environmental conditions.

In another study Freeman, this time in collaboration with Lionel G. Klikoff of Allegheny College in Pennsylvania and Harper of Brigham Young, went into the field to study the distribution of male and female plants of five wind-pollinated species of salt grass growing in the desert of northern Utah. They uncovered a significant association between water stress and gender. In patches of relatively greater water runoff, such as the bottoms of water-trapping gullies, the population tended to femaleness. At the top of these same small gullies, where water was scarcer, male bias increased.

The new evidence notwithstanding, plant scientists persisted in their reluctance to accept sexual elasticity as a natural property of dioecious species. This resistance puzzled Freeman. "You would almost have to believe there would be a lot more [sex change] in plants than animals," says Freeman, "when you think how much easier the switchover is morphologically, given the ephemeral nature of the plant's sex organs." Nonetheless, many botanists held the desert-plant studies suspect. They contended, for instance, that McArthur's cultivated plantation hardly duplicated the conditions *A. canescans* encountered in its natural desert habitat.

Refute and explain

Starting in 1978, Freeman and McArthur kept close tabs on more than 1,400 permanently marked plants. These represented seven different dioecious or subdioecious species growing in their natural environments at sites scattered throughout the deserts of Utah. Over the next three years, they turned up evidence of sex change involving anywhere from 3 percent to 30 percent of the plant populations.

In 1980, in a theoretical paper coauthored with Kimball Harper and Utah's Eric Charnov, Freeman had this to say: "We have shown that dioecious individuals with the capability of sex change will have a selective advantage over individuals lacking this ability in patchy or changing environments. We suggest that variable sexual expression may be common among species that produce unisexual flowers and that such species should possess a reproductive advantage in environments that are heterogeneous...."

The authors list about 50 dioecious species in 25 different plant families and discuss a possible mechanism by which sex change may be initiated in them. They cite several pieces of evidence suggesting that environmental stress changes the plant's ratio of cytokinin, a root-produced hormone, to gibberellin, a leaf-made product. Under water stress, for example, there is reduced availability of cytokinin, normally a promoter of femaleness, and a concomitant shift toward maleness. By the same token, circumstances that enhance cytokinin availability tend to tip the scales the other way.

What still remains to be done is to explain why stressful environments tend to push dioecious species toward maleness or less stressful environments toward femaleness. Two hypotheses have been put forth. One is Charnov's sex-allocation proposal and a further refinement of his mathematical model of size advantage. Charnov suggests that, by changing sex, an individual increases its fitness or fertility for the approaching flowering season.

The second proposition, designated variously as the resource hypothesis or the mortality hypothesis, centers on the fact that the energy cost to reproduce is much greater for females than males in some dioecious species. Thus, during stressful periods, plants may be induced to reproduce as monoecious or male individuals, thereby achieving some reproduction while reducing reproductive expenditures and enhancing the probability of survival.

According to this theory, plants change sex on the odds of an overall increase in reproductive lifespan. As Freeman puts it, "In theory one, the individual gets the payoff now; in theory two, the individual husbands resources until conditions change for the better." He has just embarked on an elaborate series of desert field trials designed to resolve whether sex-changing dioecious species play for the immediate or long-term gain—or hedge their bets and do a little bit of both.

Needless to say, the last decade has been an especially fruitful one to animal and plant biologists interested in nature's stake in hermaphroditism. But there is more to these efforts than merely the desire to tie up a biological puzzle. Practical payoffs are also in the offing. Take shrimp fisheries, for example. New practices devolving from knowledge gained about the sexual dynamics of hermaphroditic shrimp promise to ensure the quality and enhance the economic return of future catches.

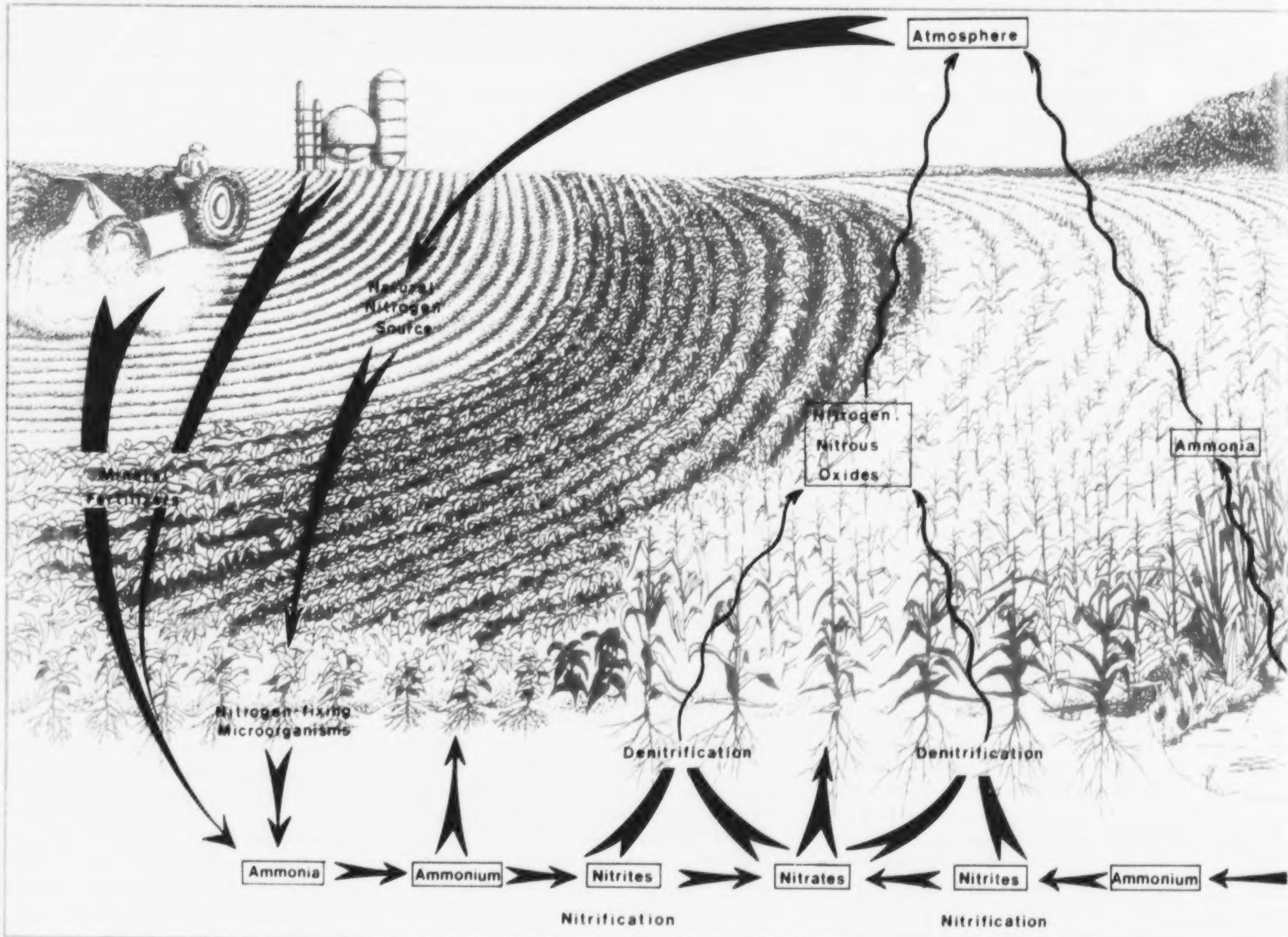
Plant specialists, meanwhile, believe the results of their sex-change studies will prove beneficial in the revegetation of ravaged desert lands and other land management situations. Already *Atriplex canescens*, a Freeman-McArthur research favorite, has been identified as one of the principal species of desert shrubs likely to be used in such enterprises. What's been learned about the relationship between sexual bent, productivity, and environment will certainly influence future plant-deployment strategies. ■

The National Science Foundation makes its contribution to the support of research discussed in this article principally through its Population Biology and Physiological Ecology Program.

The Problem of Nitrous Oxide

by Sam Iker

The increase of nitrous oxide in the atmosphere threatens the earth's protective ozone layer. The challenge is to find where it arises and how it can be controlled.


Atmosphere
Natural Nitrogen Source
Mineral Fertilizers
Nitrogen, Nitrous Oxides
Ammonia
Nitrogen-fixing Microorganisms
Denitrification
Denitrification
Ammonia
Ammonium
Nitrites
Nitrates
Nitrites
Ammonium
Nitrification
Nitrification

One of the classic environmental controversies of the early 1970s swirled around spray cans and supersonic transport planes. Scientists believed that gases called chlorofluorocarbons, used as aerosol propellants, could deplete the ozone layer in the upper atmosphere—a fragile mantle that protects the earth from the sun's damaging ultraviolet rays. Supersonic jet liners flying in the stratosphere, it was argued, could add to the destruction. The chlorofluorocarbon aerosol propellants were eventually banned and plans for an American SST were scuttled. But threats to the ozone layer have not all been laid to rest.

There are other compounds which could also affect the ozone layer if sufficient quantities were present in the upper atmosphere. One is nitrous oxide (N_2O), generated by natural processes that return gaseous nitrogen in the sea and soil to the atmosphere. It is more commonly known as the anaesthetic laughing gas.

Several years ago, scientists became concerned that vast quantities of atmospheric nitrogen were being industrially fixed (converted to nitrogen compounds), added to the soil as fertilizer, and thus introduced into the natural nitrogen cycle. If this additional burden of fixed nitrogen in the cycle generates enough extra nitrous oxide, and the gas is finding its way to the upper atmosphere, the threat to the ozone would be real. As a National Research Council report puts it, "The total amount of N_2O in the global atmosphere...is a major determinant of the rate of destruction of stratospheric ozone."

The chemical routes by which nitrous oxide could interact with ozone are well understood by scientists. Ozone is made of oxygen molecules containing three oxygen atoms (O_3); normal molecular oxygen has two (O_2). Nitrous oxide diffuses from the earth to the troposphere, where it is gradually introduced into the stratosphere by turbulent diffusion. There it is exposed to ultraviolet radiation that breaks it into its components, atmospheric nitrogen (N_2) and singlet oxygen (O). These single oxygen atoms react further with nitrous oxide to produce significant amounts of nitric oxide (NO).

During the chain of chemical alterations that follows, nitric oxide steals an atom from the ozone, combining with it to become nitrogen dioxide (NO_2) and free oxygen (O_2).

A range of estimates

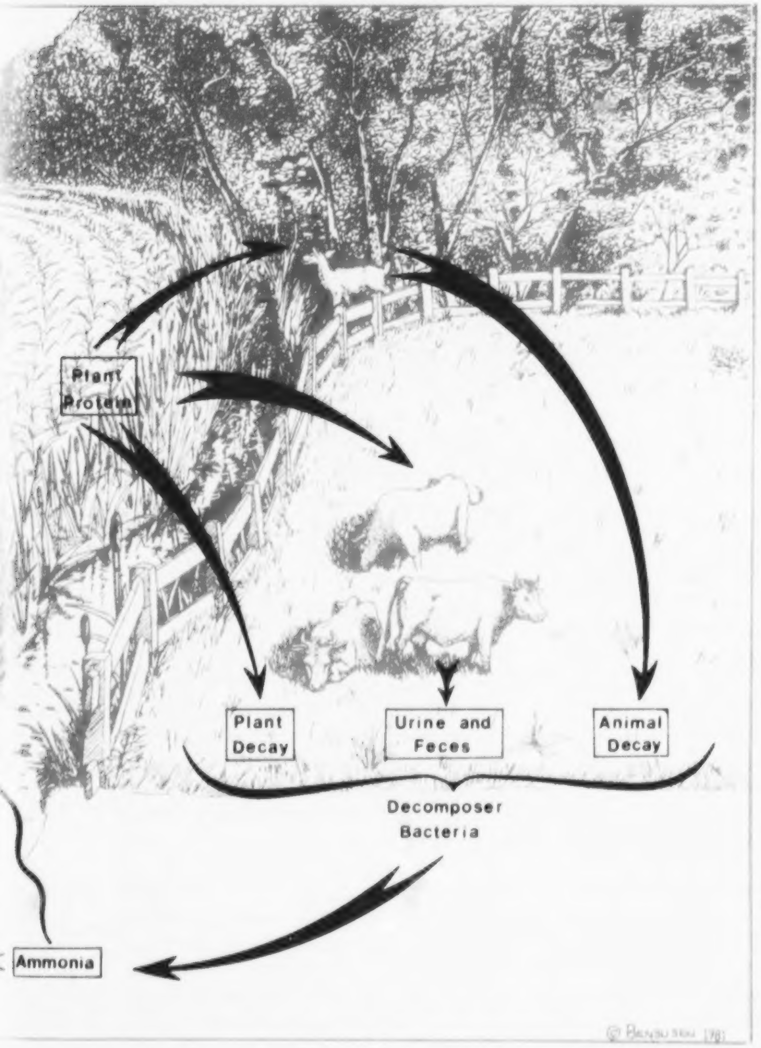
But estimates of how the increase of atmospheric nitrous oxide might affect the ozone layer have ranged widely. Scientists calculated in the early 1970s that only 3.5 percent or less of the ozone layer would be depleted. The effects might not be fully felt for hundreds of years. Others, believing that human activities were flooding the system with artificially fixed nitrogen, were convinced that the effects on the ozone layer would be far greater and felt far sooner.

Many of these researchers cited impressive numbers. The global output of inorganic nitrogen fertilizer in 1950 was 3 million tons. By 1975, the figure had soared to 45 million tons, with the United States the principal user. As a result of the rapidly growing use of such fertilizers, according to one study, as much as 20 percent of the ozone layer might be destroyed by early in the 21st century.

Nitrous oxide was a threat, all right. But how great a threat? The projections were hedged by uncertainties. What precisely happens to artificially fixed nitrogen introduced into the cycle? Was its fate different from that of naturally fixed nitrogen? What percentage of the gaseous emissions generated by natural processes is atmospheric nitrogen and what portion is nitrous oxide? How much nitrous oxide do the various parts of the land environment (both managed and unmanaged) contribute to total global emissions? How much nitrous oxide is generated by the world's oceans? And what is the lifetime of the gas in the upper atmosphere?

Over the past decade much has been learned. And much remains to be learned. In 1977, for example, some scientists downgraded the nitrous oxide threat to the ozone layer. Based on new data, investigators at the National Center for Atmospheric Research in Boulder, Colorado, altered a key chemical reaction-rate-constant in their complex computer model. The results indicated that nitrous oxide had a much smaller effect on the ozone layer than had been believed; chlorine had a greater impact.

But very recently, in only the past year or



Nitrogen cycle. Industrial, agricultural and natural conditions are involved in the chemical exchange between atmosphere and earth.

Sally Bensusen © 1981

so, new figures have been coming in. A number of chemical reaction rates have either been measured for the first time or measured more accurately than ever before with more sophisticated and sensitive instruments. Again, the new data altered the picture projected by the model. "The updated understanding we have now of atmospheric chemistry and radiation," says NCAR scientist Ralph Cicerone, "makes nitrous oxide look to be about as big a problem as we originally thought five or six years ago. The problem hasn't gone away; it could be serious."

Greenhouse effect

Additionally, research is assigning to nitrous oxide, along with carbon dioxide, a role in the production of a so-called greenhouse effect that could increase global temperatures. Both carbon dioxide and nitrous oxide intercept and absorb some of the radiation out from the earth in the infrared part of the spectrum. Infrared radiation is a principal way of giving back to space the heat the earth gets from the sun. Trapping that heat might create what could be, in effect, a worldwide greenhouse.

Oceanographer Ray Weiss of the University of California's Scripps Institution has systematically analyzed air samples from Alaska, Antarctica, and Hawaii over the past five years, along with samples collected during oceanographic expeditions in three oceans, to determine the rate of nitrous oxide increase in the atmosphere. He has also analyzed other samples dating back to 1961. These analyses show a steady rise of nitrous oxide over the past 20 years of about 0.2 percent per year. By the year 2,000, Weiss projects a 5 to 7 percent increase in the amount of atmospheric nitrous oxide over the presently measured level of 301 parts per billion. He estimates the preindustrial level of nitrous oxide was 281 to 291 parts per billion.

Weiss calculates that the present rate of increase of nitrous oxide would add about 10 to 15 percent to the effect being produced by the present rate of increase of atmospheric carbon dioxide. "That's not a trivial thing," he declares. "And when you include possible anthropogenic increases in other trace gases such as methane, you might end up multiplying the effect of nitrous dioxide by 1.5 or 2." Other studies project that a doubling of nitrous oxide in the atmosphere would have roughly the same effect as doubling the amount of carbon dioxide.

There are some other altered perceptions. Earlier studies indicated that the oceans were a major source of nitrous oxide and that the lifetime of the gas in the atmosphere was from 10 to 40 years. Weiss's measurements in ocean waters show that the seas produce comparatively little nitrous oxide. He also confirms research that puts the residence time of this gas in the atmosphere at 100 to 175 years. But if the nitrous oxide comes out of the terrestrial nitrogen cycle, where in that cycle does it happen?

Denitrification

Soil scientists have believed that the natural process called denitrification generated the excess nitrous oxide that was building up in the atmosphere. (Denitrification is the reduction of nitrates or nitrites, commonly by bacteria and usually resulting in the escape of nitrogen into the air.) They also suspected that the rapidly increasing use of inorganic nitrogen fertilizers was somehow responsible, but how this was occurring remained unclear.

There were other uncertainties, too. Was the excess nitrous oxide coming from heavily fertilized croplands? Were other parts of the land environment contributing? Again, no precise measurements had been made. One reason was the laborious and complicated task of measuring the minuscule amounts of nitrous oxide in emission samples. By the late 1970s, however, gas chromatography and other techniques made considerably refined measurements possible.

Investigators have focused on a wide range of natural and managed systems—grasslands in Colorado, wetlands and rice fields along the Gulf Coast, rich corn-growing regions in Iowa, dairy farms and croplands in Wisconsin, peat soils and alfalfa and corn fields in New York State, sugar cane fields in the Florida Everglades, upland forests in North Carolina, and river systems in Maryland and Delaware.

A better understanding has emerged of where nitrous oxide is coming from and how it is generated. "We've learned more in the last three years about the mechanisms of how gases are produced in the soil," says Iowa State University's John Bremner, "than we had for more than a decade." Some of the work confirms earlier studies; some findings provide new insights.

Denitrification requires anaerobic conditions—the presence of little or no oxygen. The scientists found that such conditions were often linked to high moisture levels. Rainfall fills the soil pores, even in well-aerated soils. Thus the oxygen supply is gradually cut off and the soil becomes anaerobic.

Large amounts of organic matter also stimulate the process. As microorganisms decompose organic matter, they consume oxygen and help to make the soil anaerobic. In such cases, temperature becomes a factor because bacterial action is stimulated by warmer conditions. Some of the scientists measured increased emissions of nitrous oxide and nitrogen gases in soils that had received heavy applications of inorganic fertilizers. Several discovered that only certain types of nitrogen fertilizers appear to generate profuse emissions.

Not surprisingly, the researchers found that denitrification occurs in spurts and varies from year to year. The highest emissions were measured during wet periods, usually spring and early summer, when heavy moisture and faster decomposition helped to deplete the oxygen in the soil. However, when the weather was dry the soil became more aerated, and when the weather was cold bacterial action slowed.

Within the framework of such findings, the researchers discovered that various biological systems, geographic locations, and soils reacted differently—and contributed differing amounts to the global flux of nitrous oxide.

Soil chemist John Duxbury of Cornell University, for instance, measured extremely high nitrous oxide emissions from the organic soils of New York State vegetable fields. "These organic soils are basically wetlands that have been drained," he explains, "and they have a large amount of nitrogen relative to what the crop can utilize. Because of the wet environment, a lot of that nitrogen is denitrified." Although such peat soils make up only about 2 percent of the world's arable land, Duxbury believes that their high nitrous oxide production makes them a "significant contributor" to the global flux.

Duxbury found that an alfalfa field, with its symbiotic nitrogen-fixing bacteria, generated far less nitrous oxide than peat soils did, but still considerably more than commercially fertilized corn fields. Somewhere in between were the emissions from a plot of corn fertilized with manure. When Duxbury applied fertilizer to alternate rows of corn, he observed ten times as much nitrous oxide coming from the fertilized rows as from the unfertilized.

Wetlands and prairies

Louisiana State University microbiologist William Patrick studied marshes and rice fields near the Gulf Coast. "What we found," he reports, "was that these natural wetlands are generally so anaerobic that denitrification largely tends to go all the way to gaseous nitrogen and not stop at the nitrous oxide phase." Patrick's results held even for fertilized rice fields. Because of the ideal conditions, according to Patrick, the denitrification cycle was "tight" or highly efficient. Consequently such wetlands seem to contribute little nitrous oxide to the atmosphere,

---

Iker has written often about environmental problems.

compared to other systems.

The same seems to hold true for the generally arid prairie lands that cover much of the West. Ecologist Robert Woodmansee of Colorado State University and his co-workers expected to find virtually no emissions from short-grass prairies. However, their measurements showed a loss of approximately a tenth of a kilogram of nitrous oxide per hectare each year during the growing season in fertilized grassland. But the year was atypically wet, Woodmansee cautions, and the readings were taken after a two-week rainy period.

Woodmansee's group also studied the effect of cattle on the system. Urine patches, which are high in ammonia, were found to generate two to three times as much nitrous oxide per unit area as the natural grasslands; they cover only a small area of any pasture, however. "The grazing animals have a significant impact," he feels, "but I'd expect it to be small in comparison to the midwest corn belt and the fertilizer additions there."

In dairy states like Wisconsin, manure is widely used to fertilize fields. And, like Duxbury, University of Wisconsin soil scientist Dennis Keeney noted a significant flow of nitrous oxide from such organically enriched soils. He also measured considerable amounts of nitrous oxide coming from drained marshlands. The nitrous oxide emissions generally coincided with periods of rainfall that helped to create anaerobic conditions. Keeney also noted high outputs of the gases as the soil thawed in the spring. The resulting rise in microbial activity could exhaust the oxygen supply. He also speculates that nitrous oxide trapped beneath frozen layers of soil might be released by the thawing.

Keeney studied one plot where minimum-tillage techniques were used—techniques that leave plant residues on the soil surface. "We're finding," he says, "that nitrous oxide pretty much follows plant residues. Before and after planting and fertilizing, we measured increased nitrous oxide emissions from the spaces between rows which were heavily littered, compared to those that were relatively free of residue."

Forests

Some of the findings surprised soil scientists. For example, prevailing theory has been that well-aerated, acid soils of upland forest regions add virtually no nitrous oxide to global emissions. Microbiologist William Caskey of the United States Department of Agriculture came up with indications to the contrary. While working with University of Georgia scientists, he studied two upland forest areas in the mountains of North Carolina. One had been recently clear-cut. The other was in a relatively natural state. It was the first long-term study of denitrification in such systems. "We're measuring the total nitrogen loss," Caskey explains, "nitrous oxide plus nitrogen." The maximum loss from the natural forest area came to 9.6 kilograms per hectare per year. The corresponding annual loss from the clear-cut site was about 25 kilograms—nearly three times as high.

In the natural forest, according to Caskey, denitrification occurred in depressions where rainwater collected and saturated the soil (making it anaerobic). The situation was different in the clear-cut area, which was thickly covered with litter from the timbering operations. Explains Caskey: "Once you cut the trees, the sun gets down directly on the soil and provides more warming. This stimulates decomposition activity." The result is an increased availability of nitrogen in the soil, and the creation of conditions that lead to denitrification as the bacteria deplete the available oxygen.

How much nitrous oxide is being generated by these forest lands? "We really don't know what the ratio of nitrous oxide to nitrogen is," says Caskey, who is planning measurements of actual emissions. But he feels the contribution of natural upland forests to the atmospheric buildup of nitrous oxide "should be considered when we construct an overall nitrogen budget for this ecosystem." In clear-cut forest areas, he adds, "the significance is going to be greater. You can establish it as a definite percentage of what's being produced by agriculture. Still, I don't think denitrification forest systems contribute that much to the global nitrous oxide flux."

Nitrification

Perhaps the most startling new results come from Iowa State University soil scientists John Bremner and Alfred Blackmer. They maintain that, based on their findings, nitrification rather than denitrification is the source of the high nitrous oxide emissions they have measured in Iowa corn fields. "When we began this work four years ago," declares Bremner, "it was accepted by everybody that nitrous oxide was produced in the soil entirely by denitrifiers. Now our work has shown us the complete opposite."

The two researchers discovered that the nitrous oxide emissions induced by fertilization of corn fields with anhydrous ammonia (a highly concentrated source of nitrogen) were more than 28 times as great as those induced with nitrate fertilizers. Most of the nitrous oxide from the ammonia-fertilized plots was produced within six weeks of fertilization, and heavier applications produced significantly higher emissions.

The process involved, according to Bremner and Blackmer, was nitrification—the bacterial conversion of ammonia to nitrite and finally to nitrate. An intermediate step in that process is the enzymatic conversion of ammonia to nitrous oxide. Nitrification of ammonia fertilizer applied to soils is usually complete within 40 days, the time span in which Bremner and Blackmer measured the heavy nitrous oxide emissions from the anhydrous ammonia-fertilized fields. Denitrification, by contrast, starts with nitrate and occurs in spurts when conditions are just right. (See "The nitrogen cycle," accompanying this article.)

To test their findings, the scientists added heavy amounts of nitrate fertilizer to the well-aerated soils typical of their study sites. There was no appreciable rise in the rather low fluxes of nitrous oxide. Even when soil samples were saturated with water to stimulate denitrification, the emissions from those containing nitrate fertilizer still fell far below those fertilized with ammonia.

Bremner agrees that "nitrification takes place in well-aerated soils. And when there's plenty of oxygen, the cycle is tight. There's little loss of nitrous oxide." But, he explains, "in an oxygen-deficient environment, the process is slower and it's a looser cycle. Thus there's an increased production—and loss—of nitrous oxide during the first part of the cycle, the oxidation of ammonium to nitrite."

Why this happens in such conditions remains uncertain. Bremner's theory is that at low oxygen levels the nitrite accumulates in the soil bacteria and causes nitrite toxicity.

"The bacteria respond to this," he believes, "by getting rid of the nitrite, by reducing it to nitrous oxide. This hasn't been proved, but it makes sense."

Bremner feels it is this finding—that nitrification generates large volumes of nitrous oxide during low-oxygen conditions—which has "led to all the confusion. It was sort of the holy writ that when there's a low oxygen supply you get denitrification. But you also get this reduction of nitrite to nitrous oxide under these conditions. And that's been hard for people to grasp because nitrification is an aerobic process."

Fertilizing rivers

The two Iowa State investigators emphasize that their results apply to soils and conditions in Iowa (and, by inference, to the Midwest corn belt). But this fact makes their findings potentially no less significant, since the greatest proportion of the nitrogen fertilizer used in the United States is in this region and the bulk of this fertilizer is in the form of ammonium or ammonium-producing compounds.

Until Bremner's work, says the University of Wisconsin's Dennis Keeney, "most of us...thought [nitrous oxide from nitrification] was probably an artifact of the laboratory. Bremner has convinced most of us that it is real and significant in the field." Among Bremner's converts, Harvard's Michael McElroy, for example, has measured profuse emissions of nitrous oxide downstream from the Blue Plains sewage plant on the Potomac River near Washington, D.C. He flatly attributes the fluxes to nitrification. McElroy explains that the treated wastes discharged by the plant are rich in ammonia. When the oxygen level of the river is high, little nitrous oxide is produced. But, says McElroy, when the river is low and oxygen-deficient (usually during warm weather months), "the bacterial population has an opportunity to grow to a point where the conversion of ammonium to nitrous oxide is very, very efficient."

The nitrogen cycle

Nitrogen (N_2) is a colorless, odorless, tasteless gas that makes up about 80 percent of the earth's atmosphere. It is critical to life—a key component of the protein molecules in plant and animal tissues. Living things, however, cannot use atmospheric nitrogen directly. The gas first has to be fixed—combined with hydrogen and oxygen and joined with other elements to form nitrogen compounds. Ultimately the nitrogen is returned to the atmosphere. The various complex pathways by which it circulates between the air, soil, water, plants, and animals are collectively termed the nitrogen cycle.

The principal mechanism of nitrogen fixation involves nitrogen-fixing bacteria that are either free living in the soil or exist in a symbiotic relationship with certain legumes such as alfalfa or soybeans. These microorganisms extract nitrogen from the air and reduce it by enzyme action to ammonia (NH_3). Or the intense energy in lightning can combine nitrogen in the air with molecular oxygen (O_2) to form nitrogen oxides (NO_X). The nitrogen oxides react with water and become nitric acid (HNO_3), which falls to the earth in rain. (See "Acid from the Sky," *Mosaic*, Volume 10, Number 4.) Additional fixation takes place in the fertilizer industry, which thereby stakes out its own location in the cycle.

Certain plants can use nitrogen in ammonia directly, but others prefer nitrogen in the form of chemical compounds known as nitrates (NO_3^-). So-called nitrifying bacteria in the soil convert the ammonia to nitrite (NO_2^-), and then to nitrate (nitrification). Plants absorb this nitrate and use it as they do ammonia to manufacture protein and other vital parts of the crop.

Meanwhile, nitrogen continually recycles in the soil. Dead plants or litter are left over from harvesting decay, along with excretions from animals that have eaten and digested plant matter. A broad group of microorganisms and fungi decompose this organic matter and convert it to ammonia. Thus the soil is recharged with usable nitrogen. Some of the ammonia is absorbed by plants and some is again converted to nitrate by nitrifying bacteria.

How does nitrogen get from the soil to the atmosphere? That process is called denitrification. In the absence of oxygen in the soil, certain microorganisms convert nitrate to nitrite through enzymatic action. Various other denitrifying bacteria reduce the nitrite to nitric oxide (NO), then to nitrous oxide (N_2O), and finally to gaseous nitrogen (N_2). The nitrogen gas diffuses through the soil pores and back into the atmosphere. Over the millenia, the various mechanisms that make up the nitrogen cycle or web have maintained a rough balance between the nitrogen in the air, water, soil, and living things. •

McElroy has come up with comparable results in his studies of the Delaware River. He concludes that "I would expect to find a similar situation in other rivers where there are large populations and sewage treatment plants. I feel very strongly that the agriculture-human-food-chain-related perturbation in the total environment is a significant factor in the atmospheric budget of nitrous oxide.

"Rivers are basically part of the same system as the agricultural system," he adds. "We are fertilizing and accelerating the productivity of specific land-based ecosystems. And we are moving nutrients from one compartment to another. In many cases we're dumping them into rivers and estuaries where the fertilization processes resume. So we're fertilizing rivers and estuaries rather than fields."

Volume estimates

Few scientists involved in nitrous oxide research are yet ready to say which of the various terrestrial systems contribute what amounts of nitrous oxide to the atmosphere. One who has attempted some preliminary estimates is Cornell scientist John Duxbury, who cautions that his projections are based on "limited data." He estimates nitrous oxide emissions from global land-based systems at about ten million metric tons a year. Of this total, he attributes approximately 40 percent to agriculture (defined in the broadest sense). Duxbury breaks down the contributions this way: cereals (corn, wheat, etc.)—12 percent; legume crops—5 percent; vegetable crops—3 to 5 percent; grasslands in humid, temperate regions—25 percent; cultivated organic soils—5 percent; and forestlands—40 percent.

These estimates raise the possibility that agriculture (and specifically the escalating use of nitrogen fertilizers) is not as responsible for the buildup of atmospheric nitrous oxide as scientists thought a few years ago. Indeed, Ray Weiss of the Scripps Institution maintains, "it has become apparent that in soils and in natural waters, nitrous oxide is produced largely by nitrification rather than by denitrification. Accordingly, the rate of increase in the anthropogenic production of nitrous oxide by agricultural activity is less directly tied to the rate of increase in the production of fixed nitrogen fertilizers than would be the case for production by denitrification."

Weiss believes that the combustion of oil and coal in power plants "comes as close as anything to explaining what we see." He attributes a "substantial fraction" of the atmospheric nitrous oxide build-up to the combustion of fossil fuels—especially to coal-fired plants that are supplying a rising proportion of the world's energy. But he concedes that he may be wrong. "It still may turn out," he admits, "that some aspect of agriculture or sewage treatment is the major source. Measurements made of these things are difficult to extrapolate to a global scale."

If combustion, agriculture, and sewage treatment are adding to the natural flux of nitrous oxide, what can be done? McElroy suggests that sewage plants convert the ammonium in treated wastes to nitrate before discharging those wastes into the water. This would theoretically eliminate nitrification and thus nitrous oxide emissions (although it also could worsen the problem of nitrate pollution). The Blue Plains treatment plant on the Potomac is doing just this, and McElroy will be studying the effects.

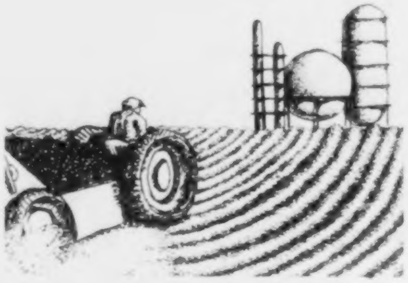
Combustion presents a more difficult problem. Current pollution controls in power plants and other combustion facilities involve particulates, sulfur dioxide, and nitrogen oxide. Nitrous oxide, on the other hand, is not even measured because it is not recognized as posing a direct air pollution threat to humans.

Agriculture is another story. Nitrification, if the process indeed is mainly responsible for producing nitrous oxide, can be modulated, a key reason Bremner and Blackmer feel their findings are significant. They have used a chemical inhibitor called nitrapyrin to slow nitrification in anhydrous ammonia-fertilized soils from a 30-to-40 day span to around 90 days, reducing the rate of nitrous oxide emissions by as much as 80 percent.

Blackmer explains that "the nitrifying organisms are a very narrow group. And we have inhibitors which essentially inhibit the activity of almost all of these nitrifiers at once." By contrast, he adds, "denitrifying bacteria are a very diverse group of organisms. At the present time we have no hope of a specific inhibitor which just inhibits denitrifying activity. Another major difference is that we know just about when nitrification is going to occur—shortly after we put nitrifying fertilizers on the soil. Denitrification, however, can happen any time."

Farmers' choice

Whether or not farmers will use commercially available nitrification inhibitors is another question. The chemicals are promoted as a way of raising crop yields by ensuring that ammonium fertilizer in the soil will be available when the plants need it in late spring. Otherwise, there is the danger that once the ammonium fertilizer converts to nitrate it can readily be leached out by heavy downpours (ammonium, unlike nitrate, does not readily wash away). However, the inhibitors add to a farmer's costs. And in drier growing areas, or when nitrogen fertilizer is applied close to planting time and the danger of leaching is reduced, there is also less of an incentive to use the chemical.



Some scientists raise the possibility, as Kenney puts it, of "moving agriculture toward an ammonium cycle as opposed to a nitrate cycle." The idea would be to employ available chemicals, or some yet-to-be discovered natural inhibitor, to delay conversion of ammonium fertilizer to nitrate until the plant can use it. At the same time, plant geneticists would select and breed plant varieties that efficiently use nutrients in the form of ammonium (many plants already do this) and produce superior yields.

Another approach to reducing gaseous nitrogen losses is simply better management. Keeney points out that minimum-tillage and no-tillage farming techniques tend to "leave more residue on the surface as a carbon source that doesn't decompose as fast." At the same time, these techniques put fertilizer closer to the soil surface. The combination creates conditions that increase the potential for denitrification. Says Keeney: "Farmers are going to have to learn to put their fertilizers somewhere else. And by cutting down on denitrification and using inhibitors on nitrification, farmers can win twice." They should also restrict the use of fertilizer to recommended rates of application and should fertilize as closely as possible to the times the plants need it, Keeney notes.

Cornell's Duxbury advocates even more drastic changes in established farming methods. "We should get away from applying fertilizer in the fall," he believes. "In many situations, it really doesn't make a lot of sense." With the encouragement of fertilizer companies, he says, farmers have traditionally laid down heavy applications during the fall because they are not as busy then. Also, the fields tend to be drier than in spring, when soggy conditions might impede the use of heavy equipment. The idea is that fertilizer will remain intact in the frozen ground over the winter and be available when the plants need it in the spring.

Duxbury and others at Cornell feel this practice is unnecessary as well as costly. He points to another approach—applying smaller amounts of fertilizer in the spring when the crop is planted and adding a side dressing of liquid fertilizer four weeks or so later when the young plants need it most for growth. Says Duxbury: "When the corn plant needs nitrogen, it needs it in a hurry. And it has to be there." Techniques like these, he thinks, are efficient, requiring less fertilizer and fewer dollars. Agronomists like Keeney, however, argue that such proposals are "sound theory [but] difficult in practice." The Cornell scientists "have not had to face the real world of farming in the Midwest, where time is at a premium in the spring."

Economics

The soil scientists working on the nitrous oxide problem are frankly as interested in improving the economics of agriculture as they are in protecting the ozone layer. And by learning how to reduce nitrous oxide emissions, they hope also to reduce overall losses of nitrogen fertilizers. For instance, Bremner and Blackmer found that as much as 8 percent of the nitrogen in anhydrous ammonia spread on Iowa corn fields can disappear into the atmosphere during nitrification. Indeed, Duxbury points out that "what we really want to know as agronomists is total nitrogen losses, not only the emissions of nitrous oxide."

As Duxbury explains it, "There are a lot of situations where fertilizer is overutilized and inefficiently managed. These situations are rapidly changing because of the steep rise in the cost of nitrogen fertilizers. Five or six years ago when it was still cheap, you couldn't convince a farmer to think about using nitrogen efficiently. He wanted to put on more than enough to make sure he got maximum crop yield. Today, with the increasing cost, nitrogen is becoming a lot more significant in terms of the overall economics of production."

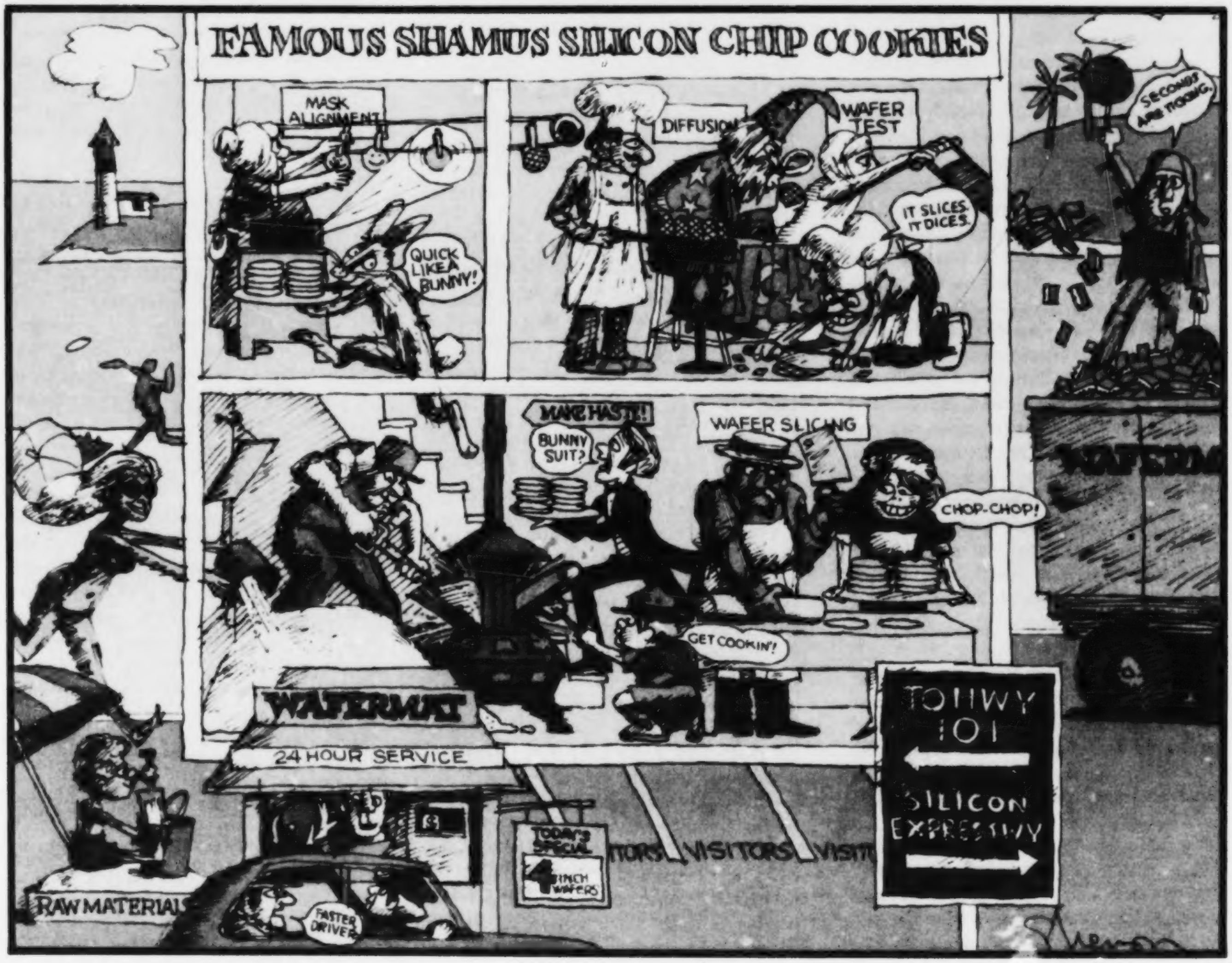
In light of escalating prices for fertilizers (most are made from natural gas) and the equally swift rise in the cost of fuel needed to apply them, agricultural researchers are seeking breakthroughs. "We must discover how to improve the efficiency of nitrogen fertilizers," declares Bremner. "It's of tremendous importance from every standpoint."

He quickly adds, echoing other soil scientists, that such discoveries would also have "great implications for the global environment." In other words, while farmers may reap the short-term benefits, everybody would benefit in the long run. •

The National Science Foundation contributes to the support of research described in this article through its Divisions of Environmental Biology and Atmospheric Sciences.

Faster Switches, Smaller Wires, Larger Chips

by Mort La Brecque


FAMOUS SHAMUS SILICON CHIP COOKIES
MASK ALIGNMENT
QUICK LIKE A BUNNY!
DIFFUSION
WAFER TEST
IT SLICES IT DICES
SECONDS ARE TICKING
MAKE HASTE!
BUNNY SUIT?
WAFER SLICING
CHOP-CHOP!
GET COOKIN'!
WAFERMAT
24 HOUR SERVICE
TODAY'S SPECIAL
4 INCH WAFERS
VISITORS
TO HWY 101
SILICON EXPRESSWAY
RAW MATERIALS
FASTER DRIVER

Very large scale integrated (VLSI) circuit technology is revolutionizing the design of computer microcircuits. Some of the mystery is coming out of the process.

Like shipwrecked Gulliver washed ashore at Lilliput, the architects and designers of tomorrow's computer microcircuits are captives of a world of the vanishingly small. And that world grows ever stranger as it diminishes in size.

This realm is bounded by a trio of simultaneous trends in VLSI (very large scale integrated) circuit technology, where it all begins: The wires and transistors that cover the silicon chips of electronic circuitry grow progressively smaller; transistor switches get faster; and chip size increases. The economic imperative is to increase electronic function per chip, so all three trends are likely to continue.

The pressures imposed on microcircuit designers are met in a variety of ways. What may be the most radical is one not yet widely taken up by industry but taught in many universities to coming generations of designers. It will thus have an impact on the way things are done in the future. This methodology was developed by Carver Mead of the California Institute of Technology and Lynn Conway of the Xerox Company's Palo Alto Research Center.

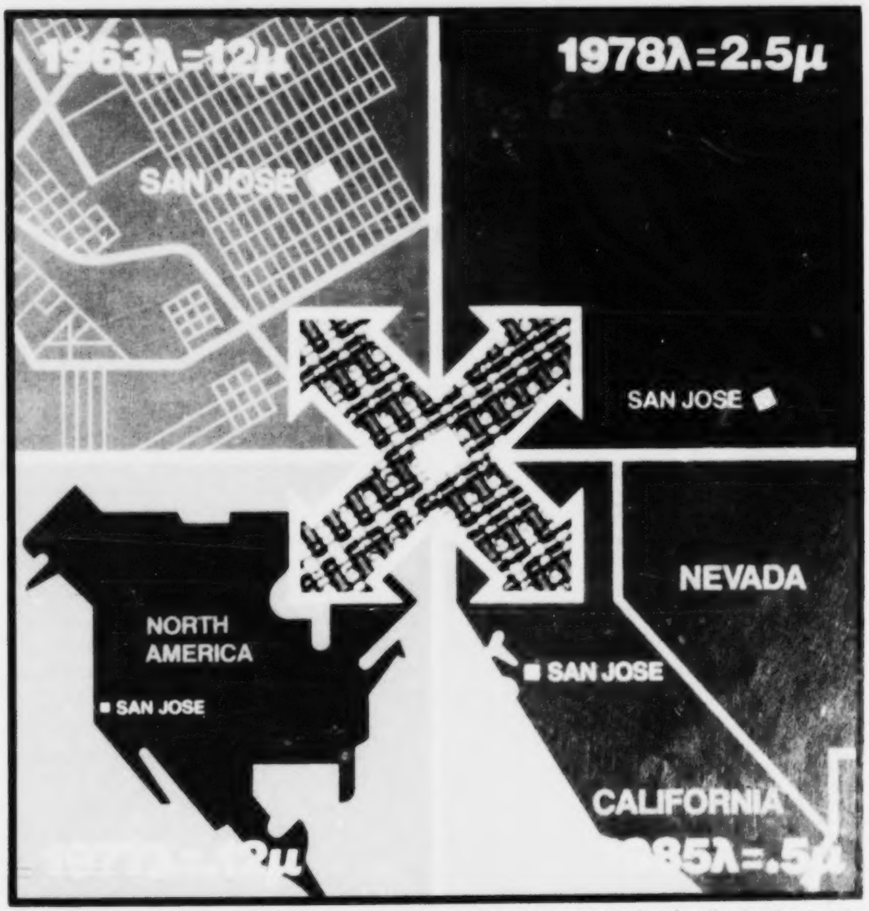
Perhaps symbolic of the volatile state of affairs in microcircuit design is a unit, that Mead named lambda, describing the distance between the individual wires in an electronic microcircuit. Individual wires in a circuit, for instance, are described as two lambda wide and the space between them is another two. But lambda, the designer's unit of distance in integrated circuits, has no fixed value. It has been made downwardly mobile to match improvements in the technology.

Shrinking lambda

To illustrate: Charles Seitz of the California Institute of Technology takes an urban street network for analogy. In the network, the distance between streets, corresponding to the separation of wire from wire, or four lambda, will be 200 meters, an average city block.

Seitz starts with the integrated circuits that appeared in 1963, when lambda was equal to 12.5 microns and chips measured a millimeter on a side. To bring that chip up to urban-network size, with its components separated by 200-meter city blocks, the scaling ratio was one to four million. The result: The millimeter-square chip becomes an area four kilometers on a side. "Most people," Seitz observes, "can find their way around in such a town."

By 1978, however, lambda had shrunk to



Circuit analogy. An analogy between integrated circuits and the complexity of a city street network shows the evolution—from 1963 to beyond 1985—of the technology. As the distance between individual wires (lambda) in a microcircuit decreases, the network ultimately could reach the density of an urban grid as large as North America on a chip the size of a postage stamp. Opposite, chips roll off a silicon foundry assembly line in a lighter technological forecast.

Circuit/roadway analogy proposed by Charles Seitz; cartoon by Mike Shenon; *VLSI Design* magazine, Palo Alto, California

2.5 microns, resulting in a scaling factor of 20 million to make the distance between wires equal to 200 meters. With chip size grown to five millimeters per side, an average chip represented an analogous area of 100 kilometers per side. That is about the equivalent of a large city like greater Los Angeles, covered by a complex network of highways.

By 1985, lambda will have shrunk further, to perhaps half a micron, Seitz says, and chip size will have grown to about 10 millimeters per side. The 200-meter scaling factor will then be 100 million and a single chip will represent, at 1,000 kilometers per side, or the states of California and Nevada populated at urban densities. Even then, says Seitz, the end will not be at hand.

If there are physical limitations to miniaturization (See "Room at the Bottom," *Mosaic*, Volume 12, Number 3), they will probably be reached only when lambda is about 0.125 micron. The scaling factor then will be 400 million. Assume that chip size has continued to increase to 20 millimeters per side, what used to be called postage-stamp size. At that point a single chip will be able to hold a detailed road map of an urban complex the size of North America.

"The complexity of something like that is just unimaginable," adds Seitz, "and the [cross-chip] communications problem will be just as bad." The reason for this, says Seitz, is that "as the number of blocks on the chip increases, so does the time required to get a signal across it. Even by 1985 standards—half-micron technology—cross-chip wire delay is far greater than the time it takes to do a simple, basic computation like an addition."

The freethinkers

In the world of very large scale integrated systems, individual chips may contain millions of transistors, and the devices and wires on them will have dimensions smaller than a wavelength of visible light. The smaller transistors become, the faster they switch on and off and the less energy they use. Energy is also saved as more transistors are put on a chip, since communication between transistors on separate chips requires tiny amplifiers and more power. Says Lynn Conway of Xerox: "If you hold the chip size constant but decrease the linear dimensions of the transistor by a factor of two, you can get four times as many transistors on the chip, all switching twice as fast. That's almost an order of magnitude improvement in what the chip can do."

Before the fruits of VLSI can be harvested, though, the problems of complexity and communication delay must be solved. A major research effort toward that end is now under way at Caltech and other institutions, including the Massachusetts Institute of Technology, the University of California at Berkeley, Stanford University, and Carnegie Mellon University. This effort is somewhat ironic, since the advances in fabrication technology that made VLSI possible happened in industry, particularly in the semiconductor sector.

Why, then, should universities be getting into VLSI design research now? Most of the industrial organizations engaged in VLSI allocate all of their available manpower to designing chips with conventional techniques, says electrical engineer Paul Penfield of MIT. "The university is an ideal setting for design research," he declares, "because we're not bound to produce a particular design on a practical time schedule and because we're encouraged to think in terms of radical new ideas."

To help bridge the distance between immediate and long-range goals, industry and the major universities doing VLSI research have built up a variety of relationships. In the most formal of these links, Caltech's Silicon Structures Project is cooperating with ten companies in the computer systems and semiconductor industries. Each company supports a full academic year of residency at Caltech for one of its employees.

Architecture or design

Designers like those in the Caltech program can help overcome VLSI design complexity, but it takes systems architects to tackle the communications problems inherent in microminiaturization. Paul Penfield draws the distinction: Design, he notes, is the series of steps needed to lay out a circuit with enough detail to produce masks from which a chip can be fabricated; architecture defines the chip's function and partitions it roughly among several large blocks of circuits.

Systems architects, it is clear, must give designers new kinds of systems to work on if cross-chip communications are to be kept from fundamentally limiting VLSI technology. The problem, Charles Seitz points out, is not just the time delay in the wires themselves. "Wires store charge," he says, "and it takes additional time for driving circuitry to apply or remove it. The act of charging or discharging the wires takes energy as well as time, in fact most of the energy used by the chip. The wires also take up most of the chip area; transistors typically occupy only 5 percent."

The costs in time, energy, and area exacted by communications are direct results of the basic structure found almost universally in computers. Called von Neumann architecture, it was devised by the physicist and mathematician John von Neumann in 1945, when active electronics were very expensive. The architecture uses only one memory, one processor, one channel for input, and one for output. This means that the act of computation must be a single, sequential process.

Since each step in that process requires signals to traverse virtually the entire machine, the concept runs into serious difficulties in the world of VLSI. There it is the communications across the chip, not the computational elements, that are dear. The solution being developed at Caltech is an approach to architecture that creates ensembles of small processors connected in a number of regular ways. Each locality is small enough that time spent in internal communications poses no problem. External communications are sparse—only between neighboring localities, not across the entire machine.

"Just as politically it seems wise to allow a good deal of local autonomy," says Seitz, "the computer organizations we're looking at reject the notion of global or centralized control. The key to the whole thing is concurrency, breaking up a computation into independent subcomputations that can be distributed and performed independently of the others. It's like human organizations, where a job worked on intensely by one group is only loosely coupled to other organization activities. The trick to be discovered is how to orchestrate the computation for the ensemble."

Systolic arrays

Coordinating the computational efforts of independent processors is indeed a major problem. Obviously they cannot be scattered randomly across the chip. If they were, communication between them would be a bottleneck in both performance and design. Computer scientist H. T. Kung of Carnegie Mellon University and his research team are seeking a way around the problem with arrays of simple computational cells that keep the data flow local and regular.

Kung calls the arrays systolic because data flows rhythmically from the computer memory through many computational cells before it returns to the memory. Kung compares systolic flow to that of an auto assembly line where different people work on the same car at different times and many cars are assembled at the same time. But an assembly line is linear, while systolic arrays can be two-dimensional, making use of even higher degrees of parallelism. They can be either rectangular, triangular, or hexagonal; to carry out different computations, data may flow on a systolic array at multiple speeds and in multiple directions.

Systolic arrays are suitable for VLSI, Kung declares, because of their regularity. He believes that most computationally intensive tasks are inherently regular and therefore can be done with regular structures such as systolic arrays.

"In conventional computers," notes Kung, "they build one general-purpose machine and program it to do different things. We achieve performance as well as flexibility by putting the building blocks of systolic arrays together in a regular manner to fit special needs."

Although general-purpose machines will not become obsolete, says Charles Seitz, he agrees that the special-purpose machines will come into their own with the new architecture. "VLSI performance is too dependent on the physical arrangement of things," he adds. "People are going to make chips that do exactly what they want them to do."

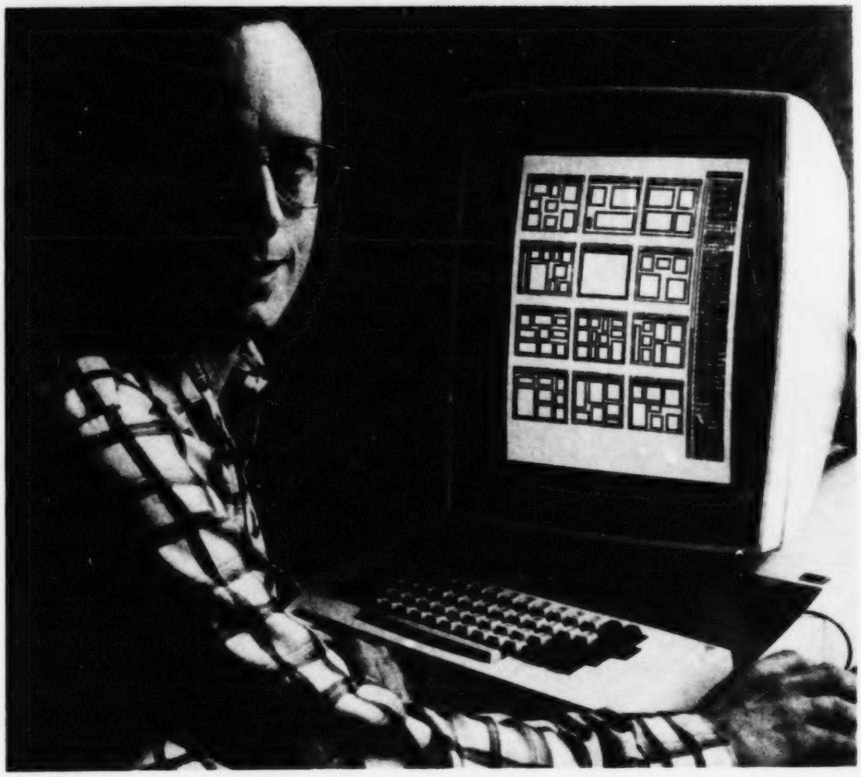
Design

Architecture is only one focus of research activity; chip design presents an even more critical problem. Designer productivity—how many circuits can be laid out in the mask per unit of time—has plummeted dramatically as VLSI technology has advanced: It can take a team tens of man-years to design the most densely packed chips (so far Hewlett Packard's half-million transistor chip is the densest), and design cost is approaching a million dollars per chip. Since chips carrying many more transistors are on the horizon, efficient design has become a top priority.

The design process is essentially a cascade of steps; in each succeeding step, the previous level of description of the circuit is made more detailed. The first level is the most abstract. The architect determines the functions of the digital system under design and draws a block diagram in which these functions are divided into subunits.

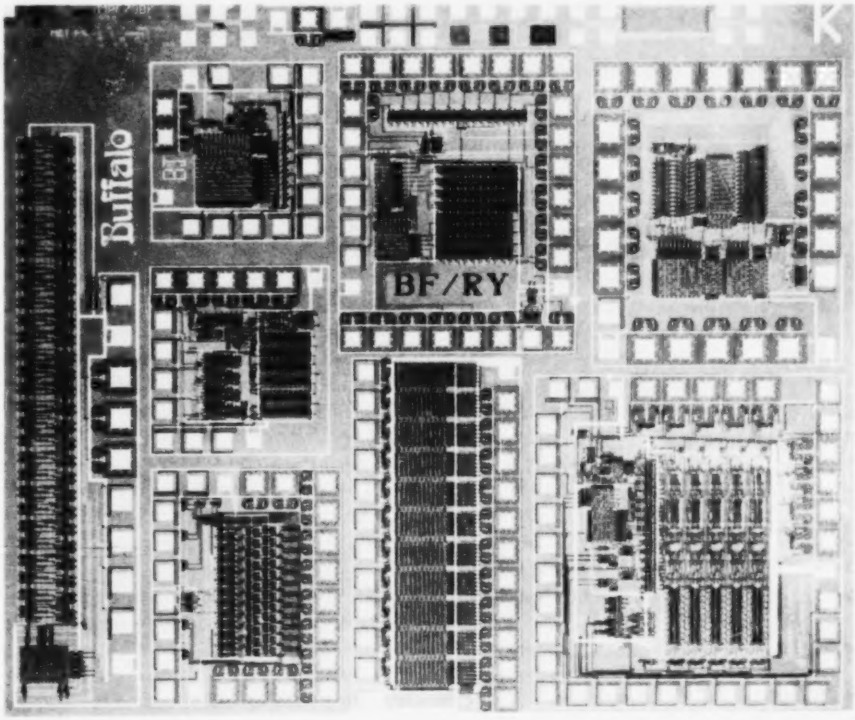
As currently practiced in the semiconductor industry, the block diagram is then given to a logic designer, who transforms it into logic-gate specifications. These go to a circuit designer, who produces electronic circuit diagrams. A layout designer works out the placement of the circuit diagrams on the silicon chip, and a draftsman painstakingly draws this symbolic layout by hand. This lowest level of design provides the details from which a working chip can be made.

This handoff process is further complicated by the fact that it is iterative. Bugs are often found and corrected, or changes made to improve performance, at a particular level of description. Steps in the design process must then be retraced to the higher levels. A round of modifications through all levels may even be necessary if the chip, when



Beginning and end. Alan Bell of the Xerox Palo Alto Research Center merges projects from 124 participating designers. One of the first multiproject chips (below) was designed by students and faculty researchers at Stanford University.

Lynn Conway/Xerox Palo Alto Research Center



manufactured, does not perform as expected.

Tall, thin designers

Must design require so many expert craftsmen, take so much time, and cost so much money? Carver Mead of Caltech and Xerox's Lynn Conway say no. They believe that VLSI design can be made so easy that one person working alone, with no previous design experience, can quickly design a chip after a one-semester course. And graduate students with backgrounds in computer architecture and systems are now learning this VLSI design methodology, outlined in a textbook coauthored by Mead and Conway, in courses at more than a hundred universities.

What these two contributed to leaner VLSI design was the realization that it could be done without knowledge of semiconductor technology, the realm of electrical engineers. What is necessary, however, is the organizational skill of computer scientists, to whom issues of complexity have become familiar ground.

Mead and Conway take a highly structured approach to designing digital systems. It is analogous to that of structured programming languages, which try to reduce errors by putting constraints on software programmers. In the absence of a sure knowledge of device physics among fledgling designers, Mead and Conway offer protection against serious design error by providing rules and many design examples.

The result is the birth of what Carver Mead has called tall, thin designers. Although not expert at any one level of description, they can successfully navigate through all the levels. Generalists rather than specialists, they are not limited to local perspectives of design problems. Nor will they encounter communications problems when changes made at one level must be transferred to other levels.

For all its effectiveness in educating designers, the Mead-Conway approach has yet to be widely adopted by the semiconductor industry. The designs that result so rapidly are not optimal—they get fewer chips per wafer and they don't squeeze the last square micron out of the silicon—and the prevailing industry philosophy is that silicon area filled is more important than design time taken. The more circuits packed per unit area, the smaller the chip can be made. Since larger chips are more subject to defects than small ones, the yield of working chips designed with the Mead-Conway method would be unacceptably low.

Lynn Conway points out, however, that industry has been exclusively interested in making products, not in exploratory design. "When you're prototyping," she says, "you don't want to invest a huge amount of design time just to confirm that the chip is going to work a certain way. Once you have a working chip, you can then optimize it if necessary. The methods we teach are intended to be starting points for rapid, exploratory design."

Genesis

The Mead-Conway approach evolved from design courses Carver Mead started at Caltech in 1970. Students were instructed in simplified versions of then-current industry practices and asked to design their own small circuits. Mead used his industry connections, Lynn Conway recalls, to have the student designs implemented. "Design is learned by doing," she declares. "It's not a spectator sport."

Conway began a research collaboration with Mead in the mid-1970s. They worked to create, debug, and document a new, simplified VLSI design methodology that the average digital designer could easily learn and apply. Their efforts led to publication of the Mead-Conway text. Conway then put together the plan for a VLSI design course for the wider university community. She taught a prototype of that course at MIT in 1978.

Student designs coming out of the MIT prototype were sent to the Xerox research center over an electronic communications network, funded by the Defense Department's Advanced Research Projects Agency to link computer science research centers throughout the country. At Xerox the designs were merged and converted to mask specifications for a multiproject chip set that incorporated a number of distinct devices on each of several chips. The masks and chips were made by companies nearby in what has come to be called Silicon Valley.

The advantages of multiproject chips were soon obvious. Students received packaged project chips six weeks after designs were sent out. This was far quicker than the usual industry turnaround and even more remarkable for new designers at a remote location. Moreover, the $20,000 cost of mask making and wafer fabrication was divided among 19 projects.

In 1979, Mead-Conway courses were given at 12 universities, all on the ARPA network. The schools submitted a total of 82 design projects on the first large-scale multiproject chip. An even larger effort with more than a hundred projects was run in the spring of 1980.

The Xerox research center operated both designs as demonstration projects but could not continue the service indefinitely. However, ARPA began large-scale funding of university research activities in VLSI design at about that time; it agreed to support the multiproject-chips project on a long-term basis. The University of Southern California's Information Sciences Institute now runs the service for universities receiving VLSI research funds through ARPA. Called MOSIS, for metal oxide semiconductor implementation system, it has already made several runs.

Of the hundreds of designs carried out so far on the multiproject chips, perhaps the most striking is that for a computer-graphics system called the geometry engine. The system has a half-million transistors, as does Hewlett-Packard's latest microprocessor which took a small team about 18 months to design. The engine was devised in a short time by one man, James Clark, a Stanford University electrical engineer who had previously designed digital computer systems but never an integrated circuit. A short Mead-Conway course transformed him into a VLSI designer.

The geometry engine can produce two-dimensional polygons, lines, and characters. When combined with a method for storing its geometric elements and converting them to images on a screen, the engine is a general-purpose computer-graphics system. As such, it can be an aid in designing such things as integrated circuits. Thus the geometry engine, itself a VLSI device, may help designers produce more of the same. Clark describes the engine as lying "somewhere between a von Neumann architecture and a systolic array." Each geometry engine is a von Neumann machine, he notes, "but the system of 12 engines is systolic, being organized into a pipeline."

Toward design automation

Current industry chip design relies heavily on graphics systems, with the designer seated at an image display unit and asking the computer to generate a wire, for example, or a rectangle. However, many university people think that a graphics-based style becomes awkward in the world of VLSI. "The level of complexity is such," says MIT's Paul Penfield, "that you need the higher expressiveness a programming language can give you."

This programming-based style, under development at the universities, relies on the computer's help in designing chips just as the graphics style favored in industry does. In fact, a small band of visionaries would like to see the computer do the entire design job itself.

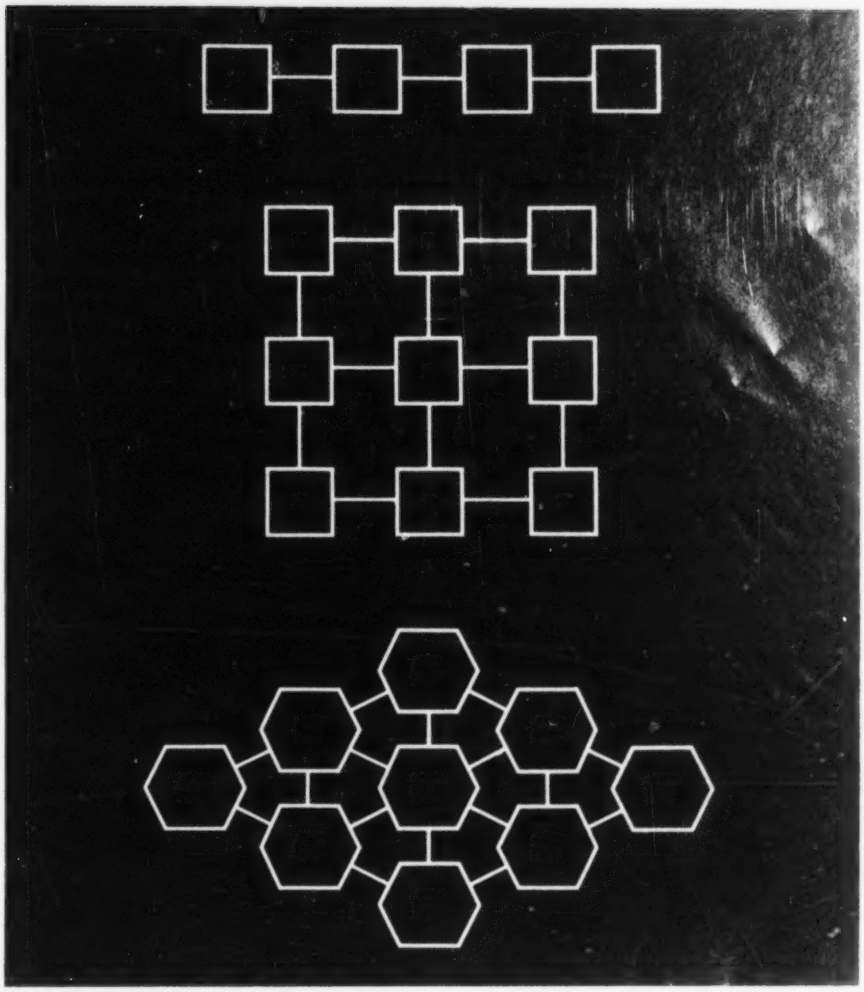
In true design automation, notes Daniel Siewiorek of Carnegie-Mellon University, the computer synthesizes alternative structures, evaluates them, and chooses the best one. Siewiorek is part of an ambitious Carnegie-Mellon effort to automate logic design, beginning with a behavioral description and ending with a set of masks for a chip that achieves the behavior desired.

Only parts of the system have been completed, but they produce in a few days designs that compare rather well with small parts of hand-tailored computers of five or so years ago. And they need only 20 to 30 percent more chip area. Siewiorek expects that systems makers, which put chips in large computers, will ultimately use design automation. With their products in a constant state of evolution and with a small product run over which to spread design cost, systems houses, unlike semiconductor manufacturers, place a high premium on saving design time and a low premium on optimizing design.

The design automation work of David Johanssen, a Caltech graduate student, has already produced complex integrated circuits with a program he calls a silicon compiler. Software compilers are programs that translate software written in high-level languages like Fortran into machine-readable form. Johanssen's program translates a high-level circuit description that merely specifies the circuit's functions into a detailed layout.

Called bristle blocks because of the rectangular cells covered with wires that form the building blocks of the circuitry, the Johanssen compilers can help design only one class of microprocessor. They would be hopelessly inefficient if required to express a wide range of functions. But within their narrow range, bristle blocks design in five to ten minutes a chip layout that would ordinarily take three man-years.

Not everyone is so enthusiastic about the complete automation of design. Some, including electrical engineer Steve Director of Carnegie-Mellon, argue that the design activity must be a partnership between the designer and the computer. "Computers plod along, doing the same calculation over and over but in general do not make any judgment about results," Director says. "People have taste and can draw conclusions from a small amount of data. Since we haven't yet figured out a systematic method for gaining such insight, we can't write a program for the computer to do it." In his own design work, Director uses a computer to help him spot trends. It can select the best way to start changing circuit characteristics and can change them until the designer sees improvement.


Systolic arrays. Designer H. T. Kung calls his arrays systolic because data flow from the computer memory in a rhythmic fashion. The system can be linear or two dimensional.

H. T. Kung

Computer-aided design is being used in two projects based on concepts in artificial intelligence; their long-term goal is to build machines clever enough to do tasks normally performed only by people. Expert systems, as they are called, consist of both a knowledge base, data that represent the system's expertise, and an inference engine that uses the knowledge to suggest courses of action. The systems also have an explanation facility: They tell users how they arrived at a particular conclusion. (See "Programmed to Think," *Mosaic*, Volume 11, Number 5.)

At Stanford University, work is just starting on Palladio, an expert system to design VLSI circuits. "Much of our activity," says Palladio codirector Harold Brown, "is in trying to discover and codify the expertise of circuit designers so that it can be used by a computer. It takes a lot of discussion to learn what a designer means when he says 'I did it this way because it seemed right.'"

Further along is a similar design project at MIT's artificial intelligence laboratory. There Gerald Sussman is applying to his expert system some fundamental research on how engineers think. He expects the complex VLSI systems produced at MIT to be useful in artificial intelligence work itself. With the aid of a program they developed, Sussman and his colleagues have already designed, in one month, a microprocessor with ten thousand transistors.

Debugging

Functional testing of a chip is the last step in circuit design; the design community is just beginning to give it attention. Debugging can be done two ways, either by simulating the function of the circuit or by verifying it with a proof that it is correct. "The difference between the two," notes computer scientist Carlo Séquin of Berkeley, "is the difference between driving a car around the block to see whether it runs and taking a lawyer's affidavit based on an inspection of the assembly line."

Verification, although a formal technique, is only as good as the proofs used to do it, and the proofs developed so far are inadequate, says Séquin. Functional testing at Berkeley relies on simulation, a program that, given a particular design, replicates the behavior induced in the chip.

In the late 1960s, Berkeley developed a program called SPICE, for simulation program with integrated circuit emphasis, to simulate analog circuitry. To achieve its high degree of accuracy, SPICE simulates every circuit in every computer cycle. This makes it too slow for the hundreds of thousands or millions of devices on a VLSI circuit. A new Berkeley program called SPLICE, patterned on logic-level simulation, traces only circuits that change from cycle to cycle. It is two or three orders of magnitude faster than SPICE, almost as accurate, and should permit VLSI-circuit-level simulation.

At Carnegie Mellon, computer scientist Robert Sproull is writing programs to simulate every design level, hanging all the simulations on a common framework. Debugging should be integrated into the design process, says Sproull, beginning at the highest levels and proceeding down to the chip itself. In this approach to debugging, the silicon represents only the most accurate simulation of the design.

But getting that chip may be fraught with its own difficulties. If the designer is with a university, a computer systems maker, or an instrument company, the fabrication process is normally done by an outside organization; only semiconductor companies ordinarily have their own chip-manufacturing facilities. The many complex steps in fabricating a mask from a design file, and a wafer from a mask, are taken at independent companies. But the service at these companies can sometimes be intolerably slow, the costs substantial, and the results unpredictable, primarily because of an absence of widely accepted standards.

The idea of a silicon foundry, proposed by Caltech's Carver Mead, would overcome these fabrication problems. In addition to having a well-publicized set of standards, the foundry could offer a truly automated service using the VLSI implementation methods developed by Conway at Xerox. Designs that meet the standards would be accepted over electronic links such as telephone wires, and chips would be returned by mail. No knowledge of fabrication or discussion with the foundry would be required of the designer.

There have been plants offering silicon-foundry kinds of services, says Douglas Fairbairn, but the first to meet Mead's definition, Fairbairn declares, will be his own company, VLSI Technology, which he plans to open in the spring of 1982. He expects that most clients will be designers working on custom VLSI circuits and they will seek prototyping service, since production volume runs are handled well by current manufacturers. "The silicon foundry should make fabrication as trouble free," he predicts, "as the Mead-Conway methodology has made design."

Experimenter-theoreticians

The million-transistor chips that roll off the foundry assembly line in the 1990s will doubtless be dazzling technological artifacts. But they may not be the most significant contribution of VLSI design research. Traditionally, notes Carver Mead, there has been a division in computer science that separates the experimenters, the engineers who design chips and systems, and the theoreticians. This bifurcation, he says, has led to stagnation in the science.

The difficulties in designing VLSI chips, Mead adds, are in making it clear that a solid understanding of computation is needed and in forcing theoreticians to face computation as a physical process. It was easy for them to avoid doing so until the advent of the integrated circuit. Before then, each part of a circuit was in a different medium and worked in a different way. Logic, for example, went on in vacuum tubes and memory in magnetic cores.

"All of computation in an integrated circuit is in silicon," notes Mead, "so we can think about generalities without having to face big discontinuities in technology. Theory has to begin with an environment that lets you understand simple things."

Silicon offered the medium for that understanding when it was introduced in 1968 and when electronic integration began on a small scale, computational theory began to stretch to accommodate it. Very large scale integration, says Mead, at last promises to move computer science farther along that road in a direction the physical sciences have long traveled. •

The National Science Foundation contributes to the support of research discussed in this article through its Divisions of Mathematical and Computer Sciences and Electrical, Computer, and Systems Engineering.

Before Pangea

I have just got around to reading the March/April 1981 issue of *Mosaic*, Volume 12, Number 2. An extremely interesting edition, however, I was disappointed not to see any reference to our work in Alaska in the article by Henry Simmons on "Before Pangea." There is a long section in this article concerning the paleomagnetic work done in Wrangellia and south-east Alaska by both the United States Geological Survey workers and by Rob Van der Voo at University of Michigan, but there is no mention of the work by myself and a graduate student, D. R. Packer. We both published extensively on the topics covered in the article, starting in 1972, some five years before any of the data quoted was published, and had drawn similar conclusions.

Having had my say on the lack of recognition of our work, I would like to congratulate you on *Mosaic*. I have found every issue to be of great interest, well put together, and very informative.

David B. Stone
University of Alaska

Friendly giant

In your recent article, "A Friendly Giant," *Mosaic*, Volume 12, Number 5, there is an illustration of an edifice which represents the achievements within the field of finite group theory. The short list of names lettered on the walls should not be interpreted as an official statement of who has made the most important contributions to finite group theory. Only a detailed account can give a fair and accurate idea of how the works of many individuals have influenced the development of this complex field.

Robert L. Griess, Jr.
University of Michigan

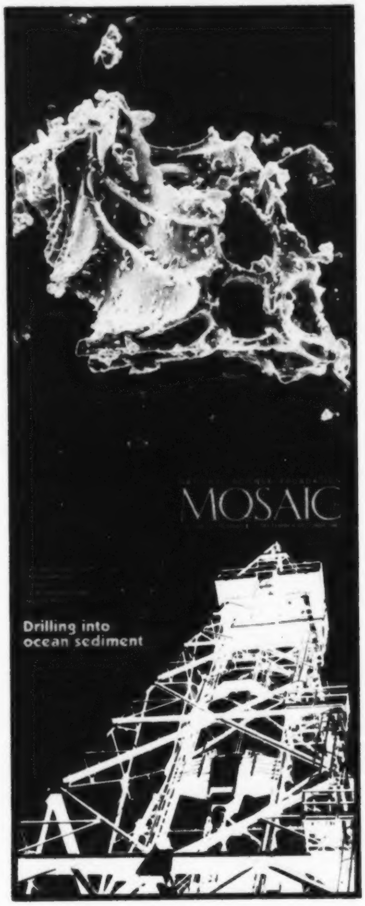
etc.

The 8.2-magnitude earthquake that leveled T'ang-shan, China, ("Earthquake Engineers Let it Slide," *Mosaic*, Volume 12, Number 6) took place on July 28, 1976.

give a serious gift
to a serious friend

or give one
to yourself

MOSAIC



Discover that science can be both readable *and* authentic.

Mosaic is a unique science magazine. It is written and edited by professional science journalists, authenticated by the scientists who know each field of research best. You don't have to be an "expert" to read it; you don't have to be a neophyte to learn from it.

Published six times a year by the National Science Foundation, *Mosaic* articles draw on the richest possible pool of research in the natural and behavioral sciences. *Mosaic* doesn't report "breakthroughs," but rather outlines the frontiers of research, in and across scientific disciplines, from which tomorrow's advances will come.

To start a subscription to *Mosaic*, for yourself, a colleague or a friend, **use the subscription card at the front of this issue.**

National Science Foundation
Washington, D.C. 20550

Official Business
PENALTY FOR PRIVATE USE, $300

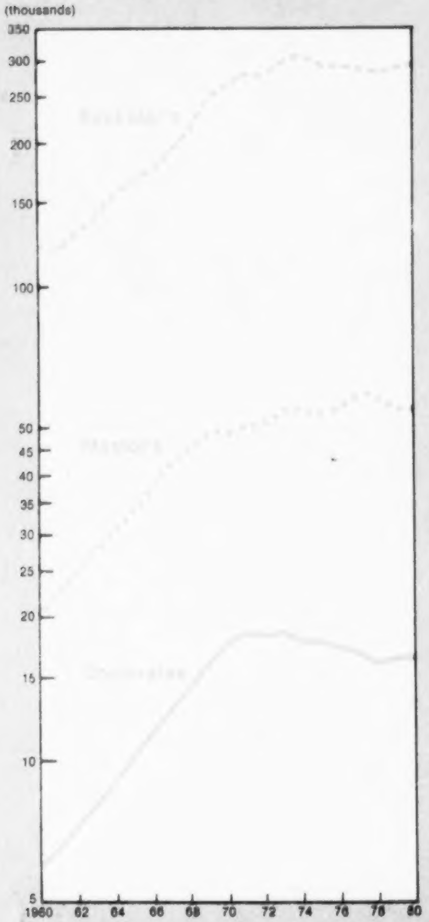
Postage and Fees Paid
National Science Foundation

BULK RATE
U.S. POSTAGE PAID
PERMIT NO. G69
WASHINGTON, D.C.

Trends in Science and Engineering Degrees

The number of science and engineering degrees awarded annually is now below the levels reached in the early to mid-1970s. This recent decline follows a long uptrend that began in the mid-1950s. Bachelor's and first-professional degrees peaked in 1974 at 305,000 and have been near 290,000 in the last five years. Master's degrees reached a high point of 56,700 in 1977 but in three years declined to 54,400. Doctorate degrees topped out at 18,600 in 1973 but have been below 17,000 in the last four years.

Science/engineering degrees by level of degree: 1960-80



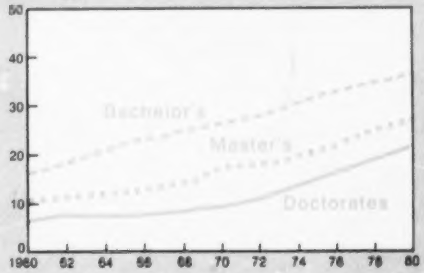
Sources: National Science Foundation and National Center for Education Statistics

The science and engineering share of all degrees fell in the last decade. About 60 percent of all doctorates until 1969, science and engineering fields now account for only half. During this same period the science and engineering share of master's degrees fell from 25 percent to 18 percent, and first-level science and engineering degrees declined from 32 percent to 29 percent.

There were substantial shifts in the relative proportions of science and engineering degrees by field, particularly at the baccalaureate and doctorate levels. Among doctorates there were marked declines in the physical sciences and engineering but increases for the social sciences. For bachelor's degrees the most notable changes were the decline in the social sciences and the remarkable turnaround in engineering from a 1976 low.

Women have earned an increasingly larger share of science and engineering degrees at all levels since 1950. In 1980 females received 36 percent of the first-level degrees, 27 percent of the master's degrees, and 21 percent of the doctorates. In 1970 these shares were 26 percent, 17 percent, and 9 percent, respectively.

Percentage of science/engineering degrees conferred on women, by level of degree 1960-80



Demographic factors make some observers believe that the leveling off in the production of science and engineering graduates since 1974 is but a prelude to further declines expected to begin in the mid-1980s. However, other factors affect degree production, as described below.

An analysis

A variety of factors combined to sustain a remarkable long-term growth from 81,000 science and engineering bachelor's degrees in 1955 to 305,000 in 1974. The Soviet Sputnik in October 1957 influenced the passage of the National Defense Education Act of 1958, which authorized direct low-interest student loans, graduate fellowships, and several forms of aid to institutions. The Higher Education Act of 1965 initiated major student assistance programs including work-study arrangements, educational opportunity grants, and insured loans. Also in the mid-1960s, the first wave of the postwar baby-boom generation reached college age.

In the 1974-80 period, bachelor's degrees in science and engineering declined about 4 percent to 292,000. During this period, the availability of student aid increased, total enrollments in higher education continued to go up, and the size of the traditional college-age population continued to grow. On the other hand, the proportion of part-time students increased, and the share of total bachelor's degrees in science and engineering declined. Also down was the number of science and engineering bachelor's degrees for every thousand persons aged 22 (the typical age of baccalaureate recipients).

These trends are affected both by changes in the relevant populations and by nondemographic factors. Beginning in the early 1960s, the population of 22-year-olds grew rapidly, which may account for almost 40 percent of the increase in the number of science and engineering baccalaureates from 1955 to 1980. The other 60 percent stems from other factors that affect students in all kinds of fields, though not necessarily to the same extent. Among those factors are changes in the value of education as perceived by young people, their parents, and employers; the availability of desired educational opportunities; and the economic capability of would-be students, including the availability of various types of student aid.

Population projections do not indicate any major changes in the number of 22-year-olds by the year 1984. Therefore, demographic factors should have little influence on the number of bachelor's degrees in the near future; changes over the next few years will primarily result from nondemographic factors. From 1984 through 1990, the population of 22-year-olds is projected to drop to a level about 17 percent below the estimated 1980 population.

13/2

NATIONAL SCIENCE FOUNDATION

MOSAIC

VOLUME 13 NUMBER 2 MARCH/APRIL 1982

LD S & T
NEWARK PUBLIC LIBRARY
FILE ROOM

Unisexual vertebrates
High-strength concrete
Chinese science
Anthropic cosmology
Plasmids

NS 1.29:

In this *Mosaic*

The oracle bone with its primitive ideographs depicted on the cover of this *Mosaic* dates to some 1,300 years before the common era (B.C.E.), in the Shang, or Yin, Dynasty of ancient China. It may be the most ancient extant record of a nova, the short-lived, flaring death of a star; the legend has been translated to say: "On the seventh day of the month...a great new star appeared in company with Antares." A similar bone inscription of the same period describes the fading of the "guest star" some two days later.

Antares is the bright red star in the constellation Scorpio (cover and inset). It is the assumption of the translator that the ideographic term, fire star, refers to Antares, as it does in later Chinese astronomy.

The short life of the astronomical event, whatever it was, makes it unlikely that it was a supernova. And in the absence of additional evidence, astronomers are reluctant to accept the Shang bone inscription as adeqnate evidence even of a nova, which would leave no detectable remnant behind. The most ancient such event for which astronomers are confident there is adequate evidence in historical Chinese records was not to occur until some 1,000 years later.

Whether or not the event was what it is said to be, whether or not the term fire star in the ideograph refers to Antares or some other bright celestial body, the Shang oracle bone has a special significance.

The time of the Shang dynasty—roughly from the eighteenth to the twelfth centuries B.C.E.—is the earliest Chinese period for which there is historical record. Beyond it is principally legend. Yet in this record out of China's earliest historic past, there are what we would consider to be scientific observations.

This is not unique, of course; at equivalent periods to the west, in India, Egypt, Babylonia, and other accepted cradles of civilization, science and mathematics were also well advanced. In fact, it is more than likely, as Joseph Needham suggests in his monumental *Science and Civilization in China*, that these early civilizations of East and West were in touch with one another; nor is there any reason to make assumptions about the direction of the flow of information.

But somewhere along the line the patterns changed, as Needham and others have documented and as Arthur Fisher reports in this *Mosaic*. Fisher's history-of-science article, "The Roots of Science in Ancient China," probes the interruption of the growth of science and technology in China: progress that had been made originally in China overcome and surpassed in the West. And the scientific revolution in the centuries that followed the Renaissance in Europe never took place in the East.

The roots of that divergence lie deep in the differences between the two worlds as they developed. Fisher's article seeks to trace those roots by exploring the work of leading historians of Chinese science, especially that of Needham and his associates. The explanations that are emerging, from sources with orgins in both East and West, help us to understand the forces that led us all to become what we are.

Warren Kornberg

MOSAIC

VOL. 13, NO. 2 MARCH/APRIL 1982

NEWARK PUBLIC LIBRARY
JUN 16 1982
U. S. GOVT. PUBS.
DEPOSITORY 316

NATIONAL SCIENCE FOUNDATION

Lewis M. Branscomb, *Chairman*
National Science Board

John B. Slaughter, *Director*
National Science Foundation

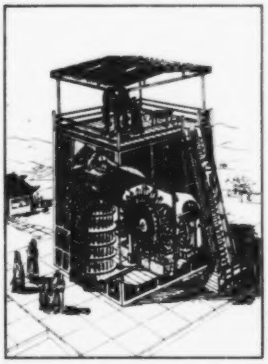
Thomas Ubois, *Acting Director*
Office of Government and
Public Programs

Warren Kornberg, Editor
David Heffernan,
Assistant to the Editor
Joyce Latham, Copy Editor

Mosaic is published six times yearly as a source of information for the general public and the scientific and educational communities served by the National Science Foundation. The Director of the National Science Foundation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this agency. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1984. Subscriptions are available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription is $12.00 per year in the United States and possessions and $15.00 all other countries. Single copies are $3.25 domestic and $4.10 foreign. Send check or money order to the U.S. Government Printing Office. Communications other than those concerning paid subscriptions should be addressed to: Editor, *Mosaic*, National Science Foundation, Washington, D.C. 20550. Telephone: 202-357-9776.

Cover. The oracle-bone inscription is reprinted with permission from *Science and Civilization in China*, Cambridge University Press; the photograph of the constellation Scorpio appears through the courtesy of O. Richard Norton, Science Graphics, Tucson, Arizona.

2



9

23

Where Males Don't Count *by Ben Patrusky* 2
Unisexual vertebrates, species that reproduce through clonal rather than sexual mechanisms, offer biologists a number of intriguing questions, not the least of which is evolutionary payoff.

The Roots of Science in Ancient China *by Arthur Fisher* 9
Chinese science once far surpassed that in the West. A 45-year-old-research effort is turning up the reasons the roles reversed.

The project 15

Redefining the Cosmos *by Henry T. Simmons* 16
Assaulting the ageless question of the nature of the universe, cosmologists have advanced a view they call anthropic. The notion bends the line where science and philosophy meet.

The riddle of large numbers 21

A Pillar of Strength *by Jane Stein* 23
High-strength concrete, used to support high-rise buildings, offers engineers an attractive alternative to structural steel.

Plasmids: More Than Genetic Debris *by William Check* 27
Ubiquitous rings of genetic material appear to researchers to have a vital role in bacterial virulence and a selfish penchant to replicate.

Plasmids in vertebrates 31

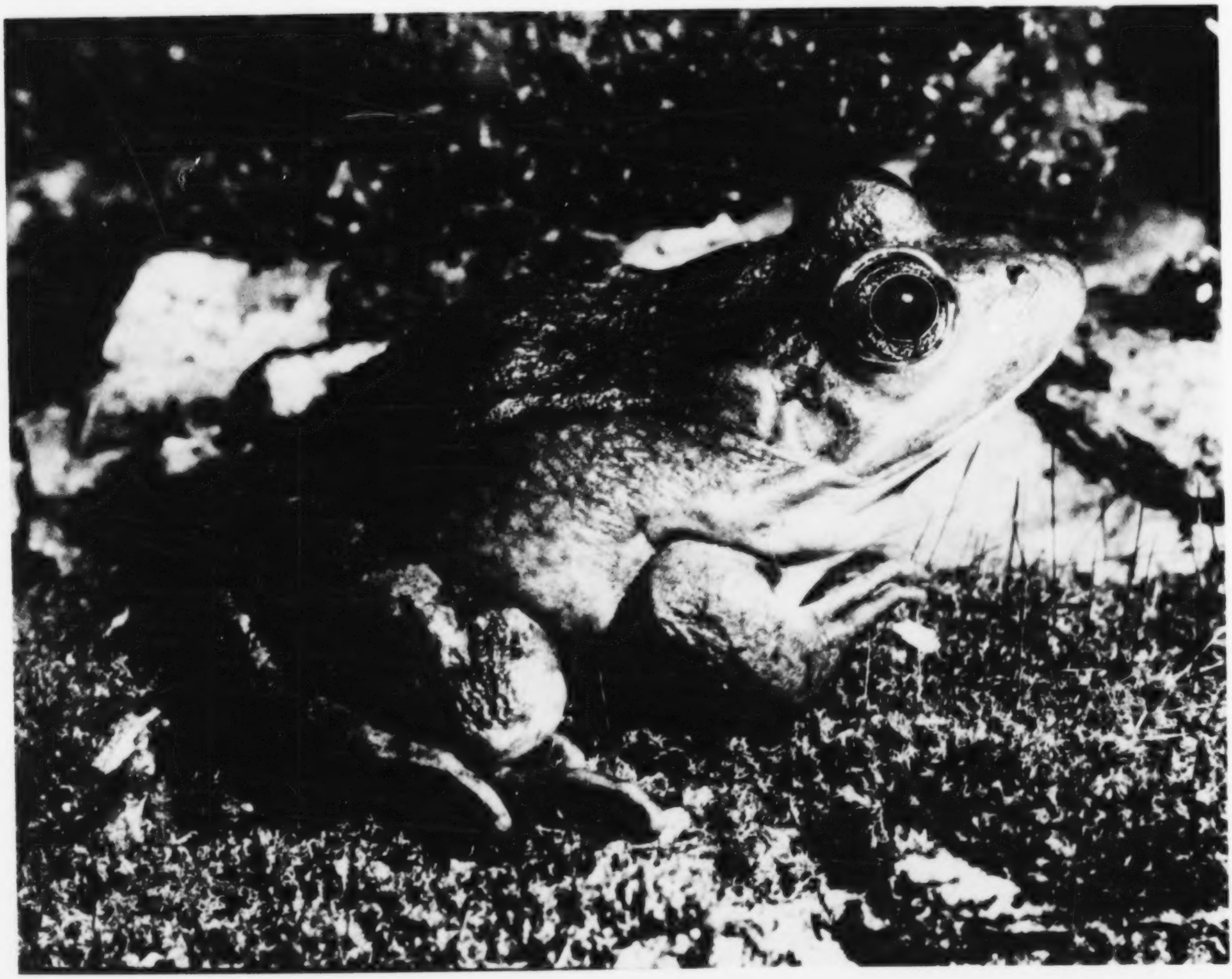
Letters, etc. Inside Back Cover

Foreign Students Are One-Fifth of Total Science/Engineering Graduate Enrollment in U.S. Doctorate-Granting Institutions Back Cover

Where Males Don't Count

by Ben Patrusky

Unisexual vertebrates, species that reproduce through clonal rather than sexual mechanisms, offer biologists a number of intriguing questions, not the least of which is evolutionary payoff.



Not very long ago, animal biologists were convinced that unisexual reproduction, the capacity to clone, existed only among single-celled organisms, plants, and a handful of invertebrate animal species. Vertebrates were presumed uniformly capable only of sexual reproduction.

Vertebrate species were of necessity, then, bisexual; each came in two kinds, male and female. From time to time reports of all-female vertebrate species surfaced, but these were invariably discounted as either wholly insignificant freaks of nature or the products of incomplete exploration.

This is no longer the case. Unisexual vertebrates are now known to exist; so far about 50 species have been discovered among fish, amphibians, and perhaps reptiles.

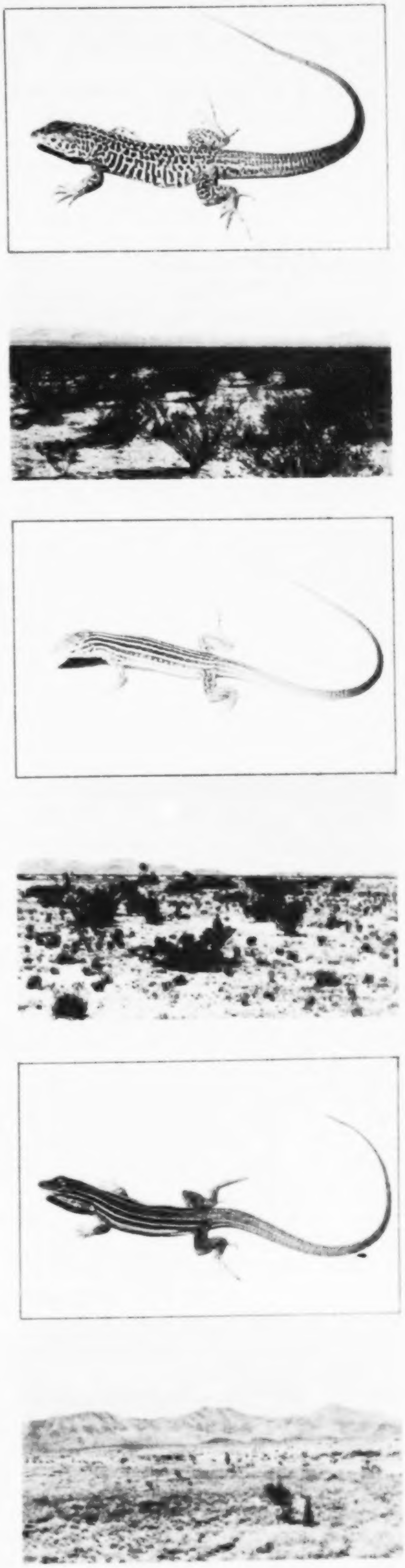
How would such phenomena come to be? The all-female vertebrate species studied in detail thus far, both in their native habitats and in the laboratory, have been found to be hybrids, products of matings between males of one bisexual species with females of another, closely related one. Generally, males strongly prefer females of their own species. But outcrosses do occur, presumably in circumstances where abrupt changes in the local ecology cause the habitats of the two bisexual species to impinge on each other. Case in point: the well-studied parthenogenetic New Mexico whiptail lizard *(Cnemidophorus neomexicanus)*.

The unisexual New Mexico whiptail arose as a hybrid between the western whiptail *(C. tigris)*, which lives primarily in the desert, and the little striped whiptail *(C. inornatus)*, a grassland species. According to Charles J. Cole, curator of the department of herpetology at the American Museum of Natural History in New York, hybridization between these lizards probably occurred in the desert-grassland transition zone where the two bisexual species were likely to cross paths.

The hybrid whiptail, as would be expected, is intermediate between its bisexual parental ancestors in many characteristics. "It follows," says Cole, "that the hybrid offspring were ecologically successful in the hybrid habitat. After all, they had genes from each of two different parental species well adapted to the habitats spanned by the [intermediate zone]."

Hybrid. The common green frog of Central Europe is a cross between an aquatic species and a terrestrial one. The germ-cell chromosomes from its terrestrial ancestor somehow don't survive in offspring. Researchers are zeroing in on the factor that enables one genotype to snip out and suppress its partner's genes.

R. W. Van Devender



Transition-zone hybridization. The parthenogenetic New Mexico whiptail lizard (center) is a cross between the western desert whiptail (top) and the striped whiptail, a grassland species. The hybrid inhabits an in-between ecosystem.

C. J. Cole

Three mechanisms

As research into unisexuality progressed, scientists recognized the existence of three distinct unisexual reproductive mechanisms: gynogenesis, hybridogenesis, and parthenogenesis. Two of these systems involve insemination of the hybrid female with sperm from a male of one of the parental bisexual species. Male participation may be required in the breeding scheme, but the male's genes either do not take hold in the offspring or have no effect beyond the first generation. Researchers often bracket the three mechanisms under the generic label clonal reproduction.

In gynogenesis, a pattern encountered in several all-female species of salamander and fish, the female is inseminated by a male, but the sperm does not fertilize the egg. There is no merger of paternal and maternal gene-carrying chromosomes; the sperm serves to stimulate division and development of the egg, which already possesses the full complement of chromosomes necessary to produce a viable organism.

With hybridogenesis, as seen in some amphibians and fish, there is fertilization. The resulting progeny manifest both maternal and paternal traits. When these offspring mature, however, the sex cells they produce will contain chromosomes from only one parental line; the ones from the other line will have been eliminated. Because only half the genes, those from just one of the ancestral species, is transmitted from generation to generation, this process is often referred to as hemiclonal or semisexual reproduction.

Biologist Robert C. Vrijenhoek of Rutgers University calls hybridogenetic and gynogenetic species sexual parasites. They "live off the sexual investment of other species at no cost to themselves," he says. "It's the other species that pays the cost."

Parthenogenesis, the third strategy, represents what is surely the purest form of unisexuality. Organisms that reproduce parthenogenetically, among them about 30 species of lizard and at least one species of snake, do not require any sperm for egg development. Having dispensed completely with the need for males, parthenogenetic animals are free to abandon the environs of their bisexual ancestors. Hybridogenetic or gynogenetic types, of course, are not.

With clonal reproduction in its various guises now unequivocably established in vertebrates, and with hybridization generally accepted as the key to its origins, the major thrust of current research centers on two main lines of inquiry:

One seeks to show how mating across species boundaries can give rise to fertile hybrids. Ordinarily, such matings, if they take, result in sterile offspring. The best known example: the mule, product of a cross between horse and donkey. The mule is almost invariably infertile, a reproductive dead end.

The second line of inquiry has to do with the ultimate fate of unisexual species. Are they doomed to more rapid extinction than bisexual species? On the assumption that unisexual populations lack genetic variation, most theoretical studies have argued that clonally reproductive species are inevitably headed for a relatively early demise. New studies suggest otherwise and open an even broader line of questioning: Why sexuality at all?

Polyploidy

Hard evidence of vertebrate unisexuality dates back at least to 1934. That year Carl and Laura Hubbs, a husband-and-wife team at the University of Michigan, published a report on a female fish population they discovered in the waters of northeastern Mexico, just below the Texas border. The species was *Poecilia formosa* (subsequently dubbed the Amazon molly). In later years, the Hubbs team came to suspect—but for a long time did not have the tools to demonstrate—that their molly was the offshoot of a cross between the marbled and sailfin mollies familiar to tropical-fish breeders.

About two decades after this initial discovery, Robert R. Miller, also of Michigan and a former Hubbs student, came across another all-female fish population—this time in the streams of northwestern Mexico, south of Arizona. This fish was a guppylike organism that bore full-term young rather than eggs; it was distantly related to the molly and belonged to the genus *Poeciliopsis*.

Since then, at least six different unisexual species of *Poeciliopsis* have been found in the waters of that region, chiefly by zoologist R. Jack Schultz of the University of Connecticut in Storrs. A former student of Miller's, Schultz has spent more than two decades studying these species; many present investigators once worked under him.

Schultz's proof of the hybrid origins of unisexual *Poeciliopsis* came in 1973. He succeeded then in what amounts to synthesizing a unisexual via direct mating experiments, initially with the bisexual fish species *P. monacha* and *P. lucida*. He has since been able to repeat these results with a success rate of about 6 percent. That is, of every hundred induced interspecies matings, about six yield clonally reproducing hybrids. It wasn't much later that Schultz uncovered another strange phenomenon among his clonal fish: polyploidy, or the presence of more than the two sets of chromosomes common to vertebrates.

Normally the body, or somatic, cells of bisexual vertebrates are diploid, having two complete sets of gene-bearing chromosomes, or genomes—one inherited from the father, the other from the mother. Egg and sperm are not diploid; they are haploid, containing just half the number of chromosomes found in the somatic cells. Fertilization—the union of a haploid sperm and a haploid egg—recreates the diploid state and gives rise to a diploid organism.

In examining the chromosomal and inheritance patterns of his unisexual species, Schultz discovered not only diploid but also triploid species, those with three rather than two sets of chromosomes. Thus, whereas somatic cells of diploid *Poeciliopsis* unisexuals had 48 chromosomes, the same number as in cells of bisexual species, the cells of triploids contained 72. By examining proteins in these fish, Robert Vrijenhoek was able to confirm these results and demonstrate the presence of genes from both sexual ancestors in the hybrids. He also turned up the protein patterns expected for diploid and triploid organisms.

Schultz found examples of triploid unisexuals among fish in the waters of northern Mexico. The morphology, chromosome and protein signatures of these fish show that they had inherited a haploid complement from each of three different bisexual ancestral species. Exactly how triploidy came to be remains a mystery, but the guess is that it resulted from a subsequent hybridization event: fertilization of an already clonally capable diploid egg.

Experimental evidence that these triploid species reproduce altogether clonally (unlike hybridogenetic *Poeciliopsis* species, which reproduce hemiclonally), came from a series of immunological trials by William Moore, a former Schultz student now at Wayne State University. He grafted tissue from mother to daughter, daughter to mother and daughter to daughter as well as mother to grand- and great granddaughter. Each readily accepted the graft from another, a sure sign that they were genetically identical. On the other hand, grafts to or from fish of one of the parental bisexual species, or from one triploid line to another, were rejected.

Finding the hybridizers

Polyploidy is not a fact of life in unisexual fish alone. Extra chromosome sets had also been found in other classes of vertebrates—reptiles and amphibians—in which unisexuality occurs. In one case, its discovery was the key to resolving a mystery that had had biologists scratching their heads for years.

The mystery involved a population of mole salamanders (genus *Ambystoma*), whose domain covers a large area of the northeastern United States and southern Canada. For a long time, it was believed that these salamanders, often referred to as the Jefferson complex, consisted of just one species. What could not be explained was why, in habitats running through the mid-portion of the salamander's range (a swath roughly 250 kilometers wide extending from Chicago to Boston), females outnumbered males by as much as 200 to 1. This occurred despite the fact that in breeding congregations of other species of *Ambystoma* there were usually far more males than females.

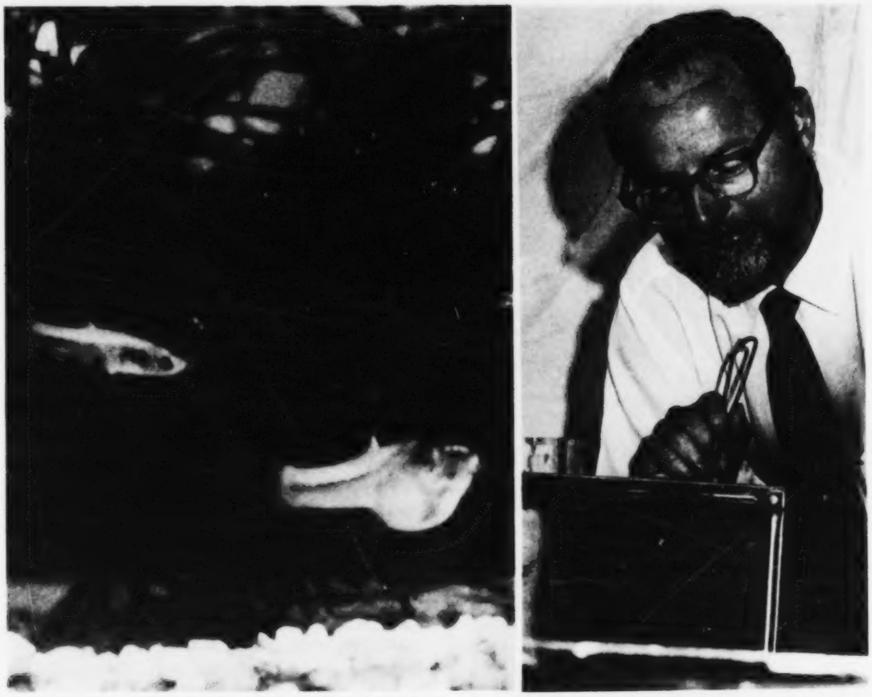
The solution came from studies in the early 1960s conducted by Thomas Uzzell, then a doctoral candidate in zoology but now curator of vertebrate biology at the Academy of Natural Sciences in Philadelphia. Uzzell was able to show that this salamander population consisted of not one but four distinct species. Two were bisexual and diploid: the large, pale, and long-legged Jefferson salamander *(Ambystoma jeffersonianum)* south of the swath and the smaller, darker, blue-spotted salamander *(A. laterale)* to the north. The other two—the silvery salamander *(A. platineum)* and Tremblay's salamander *(A. tremblayi)*—proved to be all-female triploids, hybrids resulting from interspecies mating of the Jefferson and blue-spotted types. They bred gynogenetically and inhabited the intermediate zone.

Uzzell later went on to establish the pedigree of what proved to be the first known of a number of strictly parthenogenetic vertebrates—those that reproduce unquestionably without male participation. Credit for the actual discovery falls to Ilya Darevsky of the Zoological Institute of the U.S.S.R.'s Academy of Sciences. In 1958, Darevsky published a report of female populations of Caucasian rock lizards of the genus *Lacerta*. These creatures, centered around Lake Sevan in Armenia, produced offspring despite the complete absence of males. Uzzell's aid was enlisted when debate arose as to the origins of these unisexual Lacerta, and he was able to delineate the hybridizing species that yielded these true parthenogens.

Since Darevsky's discovery, at least 100 scientific papers covering strictly partheno-

---

Patrusky is a New York-based science writer and consultant to several medical and scientific societies.



Producing unisexuals. R. Jack Schultz of the University of Connecticut succeeded in synthesizing a unisexual species of *Poeciliopsis* by crossing members of one bisexual species (left) with another.

genetic reptiles have appeared. In all, some 30 species of lizard (about 1 percent of the estimated 3,000 in the world) and one species of snake have been so identified.

Parthenogenesis

But are they all truly parthenogenetic? The likelihood is that they are, but questions remain. As Charles Cole of the American Museum of Natural History explains: "If you did an exhaustive search of the literature for solid, conclusive, unequivocal evidence that any species of lizard was parthenogenetic, you'd have a hard time finding it." Much of the evidence put forth for parthenogenesis, says Cole, turns out to be predominantly circumstantial, and "any day a male could rear its ugly head and destroy the hypothesis."

To remedy an uncomfortably ambiguous situation, Cole launched a series of studies to establish beyond doubt that species of reptiles can and do reproduce entirely clonally. The plan was to study lizards under controlled conditions. For his trials, he chose to work with species of the unisexual whiptail lizard, (Cnemidophorus) common to the Western Hemisphere. Of about 40 species in the genus, 13 are thought to reproduce parthenogenetically. Cole's research favorites were several species from the American Southwest, particularly the New Mexico and Chihuahua whiptails.

To establish parthenogenesis, Cole began with adult unisexual lizards captured in New Mexico. These he maintained in isolation from all other reptile species. In time, the females laid eggs that developed into complete organisms. But that alone didn't prove anything. "Female lizards can store sperm for months," Cole explains, and in some cases perhaps years on end. "Our field-caught lizards could have had contact with males prior to capture, which would have enabled them to produce clutch after clutch in the cage long after insemination."

To eliminate this possibility, Cole isolated the newborns at the moment of hatching. A year later, these females—definite virgins all—laid eggs that subsequently hatched out more females. The original colony is now in its seventh generation. All the offspring, from one generation to the next, have been raised without any male contact. (That all are exact genetic duplicates of each other was confirmed by Herbert C. Dessauer of Louisiana State University at New Orleans.)

But reproduction without benefit of males did not constitute absolute proof of parthenogenesis. Cole worried about hermaphroditism, the presence of functioning male and female organs in the same individual, and thus the possibility of self-fertilization.

Hermaphroditism is not unheard of in vertebrates. A few species of fish have been discovered, for example, with the capacity for self-fertilization. To check his whiptails, Cole sought the services of Laurence M. Hardy, a herpetologist and histologist at Louisiana State University's Shreveport campus. A few years earlier Hardy had discovered the phenomenon of intersexuality where, in some species of snake, adult females develop male copulatory organs but don't use them.

Cole supplied Hardy with two second-generation clutches of lab-raised, unisexual whiptails. Hardy then set about preparing tissue sections of the entire reproductive tract of each lizard, selecting a different one every ten weeks. This gave him a full growth series right up to the point of sexual maturity.

Hardy recently completed his microscopic examination of a total of 6,000 cross-sections and has reported finding nothing that resembles male tissue. For Cole, these findings no longer leave any room for doubt. "We are now absolutely convinced," he says, "that in the case of our whiptails we are dealing with true parthenogenesis."

Cole would like to see this same sort of rigorous examination applied to other suspected parthenogenetic reptiles, especially *Typhlina bramina*, the smallest snake in the world. Measuring only six inches when fully grown, *T. bramina*—commonly called the flowerpot snake—has been found in the tropic regions of Mexico, Asia, Africa, and Australia. Its identification as a likely unisexual owes much to serendipity.

Samuel B. McDowell, a zoologist at Rutgers University, Newark, had been examining the evolutionary relationships among a group of snakes that included *T. bramina*. McDowell dissected dozens of *T. bramina* specimens and found no males. He borrowed specimens from elsewhere—still no males. To date McDowell has looked at well over 200 specimens, and all are female. Other investigators, hearing of his experience, joined the search. Thus far, not one male.

What makes this discovery especially exciting, if later studies support *T. bramina*'s parthenogenetic status, is that it would mark the first unisexual identified to date among the 2,500 snake species known. "With the accumulated statistics now well in our favor," says Cole, "the odds are that *T. bramina* is a real parthenogen." The final answer, Cole says, won't come until scientists learn to rear the snake in the laboratory. Chances are that tests will point to a hybrid origin as it has for other clonal species examined thus far.

Why not sterile?

One of the crucial questions fueling current research in unisexuality deals with the mechanisms that foster fertility in hybrid species. More often than not the offshoots of interspecies mating, if indeed there are offshoots, prove to be sterile.

A widely held explanation for hybrid infertility argues for faulty interaction between chromosomes derived from related but different species. Normally, sexual organisms manufacture sex cells through a mechanism called meiosis. The process consists of chromosome pairing and two cell divisions, in the course of which the chromosome number is reduced from diploid to haploid. In the process, the side-by-side chromosomes exchange some material so that each haploid ends up with genetic material from each parental genome. In hybrids the dissimilar chromosomes don't line up and pair properly, thereby disrupting the normal meiotic machinery.

While fertile hybrids also are incapable of normal meiosis, they appear to be the beneficiaries of a strategy of circumvention. There is some evidence, says Cole, that the genome from each parental lineage, acting independently, replicates itself. Thus a diploid pair becomes a four-chromosome set and a triploid a six-chromosome set. As this cell then divides, each daughter cell acquires a full (haploid) set of each of the original genomes and is thus capable of clonal reproduction. In this process, called endoduplication, there is no genetic recombination.

Endoduplication begs a variety of questions. What makes the system work in the 50 or so unisexual vertebrate species identified so far? Is there something special about the mix of chromosomes between certain species, a cause-and-effect relationship, that automatically confers a potential for replication upon the resulting hybrid? Or, as some scientists argue, is something else required—perhaps a specific, rare mutation—in addition to a particular balance of chromosomes?

Cole's whiptail lizard may help provide part of the answer, at least where it concerns demonstrably parthenogenetic unisexuals. The idea is to see if he can produce such creatures in his laboratory. Toward this end, he has placed the males of one bisexual lizard species *(C. inornatus)* with females of another *(C. tigris)* in hopes of getting them to mate. Not sure if this combination will do the trick, he has female *C. inornatus* in another cage with male *C. tigris*. As with other sexually reproducing species, these whiptails are generally disinclined to cross species lines, so Cole's plans also call for inducing interspecies crossings by way of artificial insemination.

Cole reasons that if his laboratory matchmaking results in a parthenogenetic hybrid

Environmental adaptation. Bisexual salamanders *A. laterale* (top) and *A. jeffersonianum* (bottom), inhabiting diverse environments, can produce unisexual species that occupy a swath in between.

R. W. Van Devender

indistinguishable from a unisexual in the wild, that would all but eliminate mutation from consideration. "We would hardly expect such an exceedingly rare event as a mutation, and a very specific kind of mutation at that, to occur within our very small laboratory colony of whiptails," Cole declares.

A hungry gene

While Cole addresses the question of mutation vs. chromosome balance as it relates to parthenogenesis, Philadelphia's Thomas Uzzell is proceeding on the assumption that the answer is already in for hybridogenesis and gynogenesis. Unisexual *Poeciliopsis* can be routinely synthesized in the laboratory merely by putting the appropriate bisexual species in the same tank. This, he contends, all but clinches the argument that a unisexual's capacity to clone depends solely on molecular properties of its chromosomes. Additional independent events, such as random mutations, do not seem to count for much. Uzzell has just begun a series of studies designed to root out and isolate what he suspects to be a specific molecular fragment at the level of the genes that promotes clonal reproduction, or at least lets it happen.

Uzzell is building on a central hypothesis: that one of the various hybridizing genomes calls all the shots and thus holds the key to cloning. As a case in point, he notes that among hybridogenetic *Poeciliopsis* it is always the *P. monacha* genome that gets transmitted to the next generation, never the genes of any of the species it has hybridized with.

To Uzzell, that means there's probably a specific molecular feature in the makeup of the *P. monacha* genome that enables it to excise the other genomes before they can be incorporated into the egg. He contends that this mechanism probably operates universally in unisexuals. As Uzzell puts it: "Whatever its form, unisexuality arises from the fact that one genome does something strange to the genome from another species so that the result is inevitably a clonally reproductive hybrid."

Uzzell is chasing after that strange something through a series of complex crosses involving the common edible green frog of Central Europe, *Rana esculenta*. A hybridogenetic hybrid, the frog devolves from a cross between a large, aquatic species and a smaller, more terrestrial one. Unlike hybrid *Poeciliopsis*, the hybrid green frog comes in two sexes, but the distribution varies. In most locales, the male-to-female ratio is one-to-one. In some habitats males predominate; elsewhere females do. Most are diploids; some are triploids.

Artifact. Rutgers University biologist Robert Vrijenhoek's work with synthesized unisexual *Poeciliopsis* shows the species is prone to Siamese twinning.

Each body cell of a hybrid, diploid green frog contains a chromosome set from each parent. Its germ cells, however, almost always contain the chromosomes from only its aquatic ancestor *R. ridibunda*. Uzzell, in conjunction with Leszek Berger of Poland and Hansjurg Hotz of Zurich, is zeroing in on the genetic factor that enables one genotype to snip out and suppress its partner's genes. He and his colleagues are seeking the controlling gene fragment in naturally occurring backcrosses in which the hybridogenetic hybrid lives and mates with a member of one or another of its parent species.

Clonal adaptation

Beyond the hunt for the underlying molecular mechanism of cloning, an equally important quest is the effort to resolve the issue of unisexuals' ultimate fate. The issue comes down to this: Do clonally reproductive hybrid species face an evolutionary dead end, or do they have the adaptive wherewithal to escape rapid extinction?

As Rutgers biologist Vrijenhoek explains, "Sexual reproduction is like gambling, like playing poker. Every generation you reshuffle the deck and everyone gets a different hand. Sex results in variation and variation enables the species to keep up with environmental change. Chances are there will always be individuals with the best hand to meet the change, and so the species prevails."

Unisexuals play a different game, says Vrijenhoek. "In this game," he says, "you get a great hand right off, four aces, and you duplicate that hand and use it again and again over every subsequent generation. This game is based on the idea that conditions that establish fitness or advantage won't change. But when the environment changes, the rules of the game change. New circumstances convert what used to be a winning hand in one generation into a loser in another."

Thus by the reckoning of most theoreticians, these clonally reproductive species are doomed to a rapid demise because they lack the faculty deemed essential to long-term survival: genetic variation. Vrijenhoek, however, cites recent evidence suggesting that the origin of some extant unisexual species (*Ambystoma* salamanders and *Lacerta* lizards, for example) may date as far back as the Pleistocene. If true, that would mean that these species have been able to adapt, and Vrijenhoek thinks he has a handle on one likely mechanism: clonal variation.

Just as individuals in sexual populations differ genetically from each other, says Vrijenhoek, so may clonal lines differ from one another. A former student of the University of Connecticut's Jack Schultz, Vrijenhoek derives his proof, not surprisingly, by way of research on unisexual *Poeciliopsis*.

In one study, the Rutgers biologist turned up two distinct clonal lines among gynogenetic *Poeciliopsis* triploids imported from their home in northwestern Mexico. The individuals within each line were identical, but they differed from the individuals of other lines both genetically and behaviorally. In the laboratory, for example, one clone subsisted primarily on plant life scraped from rocks, while another browsed on floating algae, insects, and other detritus. There were also significant differences in dentition patterns.

In another study involving 37 strains of a hybridogenetic species, Vrijenhoek and Schultz, in collaboration with another Schultz student, Robert Angus, identified 18 distinct clonal lines. Genetic differences among them were evidenced by differences in protein signatures and through tissue-grafting experiments in which only individuals from the same clonal lineage fully accepted the grafts.

What these observations indicate, says Vrijenhoek, is that each hybrid event freezes a different genetic clone, each with its own specific behavioral, physiological, and ecological traits. This contradicts the predictions of rapid extinction based on the assumption that unisexual populations lacked genetic variation.

Nor, for the matter, did the short-haul theories take another potential source of variation into account: migration. Vrijenhoek cites a recent study by Schultz and Arthur Bulger that shows how a hybridogenetic *Poeciliopsis* moved 500 miles northward and acquired cold adaptiveness for its offspring by mating with a local bisexual species. In a single generation, the fish had picked up an adaptive trait from a species that may have spent thousands of years evolving it.

One other factor may be working to the evolutionary advantage of some clonal species: scientists call it heterois. That's where the hybrid, having acquired adaptive characteristics from two or more parental species, is better suited to a greater range of environments than any of its ancestors. The European green frog *Rana esculenta* is a case in point. It can hibernate either underwater, as does one parent, or on land, as does the other. Neither of the parental types is able to do both. Polyploidy adds another dimension, bestowing an even greater variety of adaptive gene combinations on the individual.

Ratcheting forward

These avenues of evolutionary flexibility notwithstanding, there seems to be an inescapable limit to long-term species survival built into clonal reproduction: a phenomenon called the ratchet mechanism, first postulated by Nobelist Herman Muller in 1964. It refers to the gradual but inevitable accumulation of genetic mutations. Since harmful mutations occur far more frequently than beneficial ones, a clone, which has no way of getting rid of deleterious genes, will continue to amass errors, ratchet forward, if you will, and eventually deteriorate.

From a recent series of studies on his unisexual *Poeciliopsis*, Vrijenhoek has gathered evidence that supports Muller's hypothesis. Vrijenhoek suggests, however, that a variety of countervailing forces are at play in *Poeciliopsis*. Both the recruitment of new clones through new hybridizations and polyploidy, in which the presence of extra gene doses serve to dilute the degrading effects of these mutations, are probably at work. If so, they would slow or prevent unisexuals from ratcheting themselves out of existence.

Short term or long

In the process of examining the ins and outs of unisexuality, scientists may finally unearth the long-sought ability to unravel one of the most tantalizing conundrums of science: Why is there sex? By the rules of classical evolution, unisexuality should be predominant. In the accepted Darwinian scheme, evolution is fueled by natural selection of adaptive traits that give the individual immediate reproductive advantage. By any reckoning, there is an advantage to unisexuality: Over the short run a unisexual will increase in number twice as fast as sexuals.

"We don't really know why clonal reproduction hasn't replaced sexual systems," says Vrijenhoek. "Certainly sex gives a clear, long-term advantage to populations, to species as a whole. But cloning provides an immediate gain. Just by this numerical edge, unisexuality should have overwhelmed sexual reproduction when it arose."

But that's hardly the case at all. According to a recent estimate by Australian biologist M.J.D. White, unisexuals represent fewer than 0.1 percent of the animal species among those groups that contain unisexuals. "We won't get answers to the why-sex question by studying sexual organisms," says Vrijenhoek. "We must look at the rarities, the freaks, the ones that don't behave by the rules. Only then can we hope to understand why evolution opted for sexual reproduction."

There is an important practical side to the unisexual studies, too. "Polyploidy and hybridization have been exploited to enhance agriculture and produce better crop plants," says Vrijenhoek. "Maybe we can learn to do the same thing with animals." He believes, for example, that unisexual *Poeciliopsis* will prove to be the perfect guinea pigs for geneticists seeking ways to manipulate fish for possible aquacultural benefit. He doubts, however, that his research favorites will ever tempt human palates. "No matter what we wind up doing with *Poeciliopsis*," he says, "they'll never taste good, not even on pizza." •

The National Science Foundation contributes to the support of the work discussed in this article through its Systematic Biology Program.

The Roots of Science In Ancient China

by Arthur Fisher

Chinese science once far surpassed that in the West. A 45-year-old-research effort is turning up the reasons the roles reversed.

A well-educated Chinese visiting England or any place in Western Europe a thousand years ago might justifiably have been disdainful of western technology. Where, for example, were the mechanical clocks? They had been telling time in China since the eighth century C.E. (common era) but were not to be invented in Europe until the fourteenth. Where were the wheelbarrows, invented in China in the third century but not to appear in the West for another thousand years? And where were magnetic compasses, gunpowder, and printing techniques—seen ultimately by historians of the western world as focal to the transformation of European society? All had previously been invented in China.

Indeed, virtually until the sixteenth and seventeenth centuries of the common era, the Chinese were far in advance of their European contemporaries in knowledge about nature and the application of that knowledge to human improvement. This is becoming increasingly unquestionable, even unassailable. Yet this fact raises questions, and modern scholars are only now coming to terms with them. What enabled the Chinese to gain so formidable a technological edge in the first place? And, more puzzling, with that edge why did they not proceed to what in the West became the next seemingly inevitable step: the leap forward into modern science that we call Europe's scientific revolution, which finally took place in the sixteenth century?

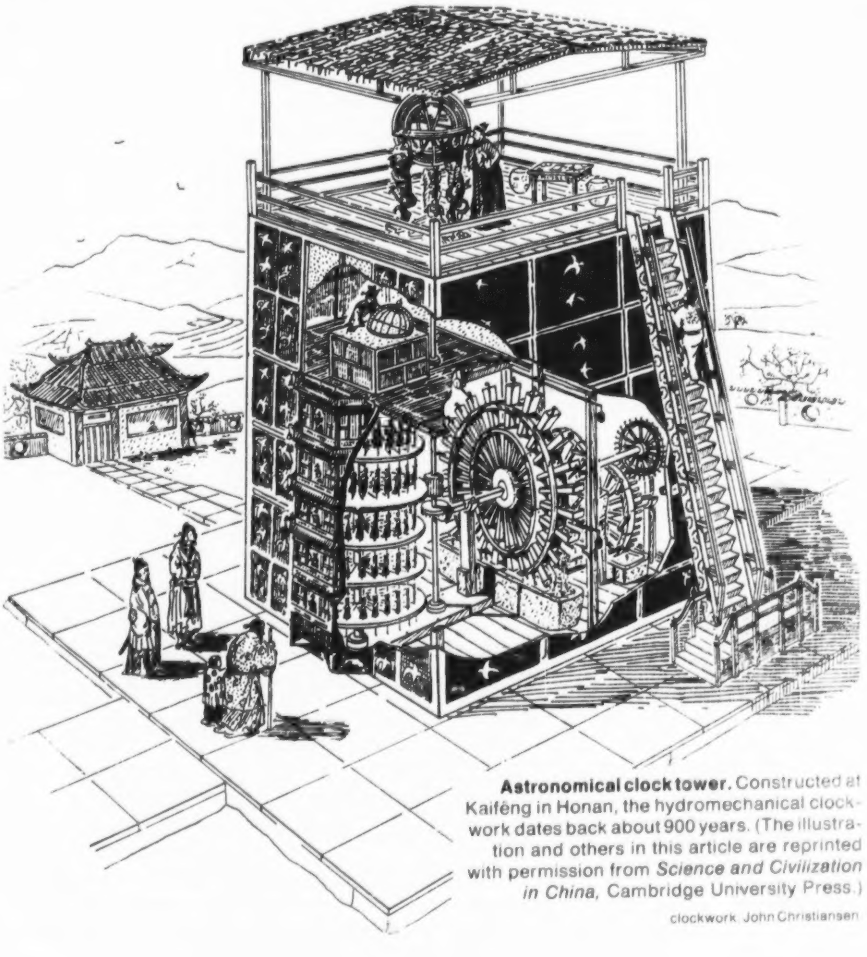
These questions are the enduring concern of an international group of sinologists and science historians. At its nucleus, as he has been for 40-odd years, is a reconstructed Cambridge University biochemist-embryologist named Joseph Needham.

Expendable

Needham, who is credited with putting the S in UNESCO (he became science director), had been sent to China during World War II to be scientific counsellor at the British Embassy in Chungking and to run the Sino-British Science Cooperation Office. He had learned Chinese from colleagues at Cambridge. "There were very few British scientists who knew Chinese," he recalls, "and a biochemist-embryologist was expendable."

"Those four years in China sealed my fate," adds Needham. "After that it was impossible to think of doing anything but a book on the history of science, technology, and medicine in Chinese culture. It was something that had not previously existed in western literature."

That was the genesis of an international exercise that is still going on, one that has



Astronomical clock tower. Constructed at Kaifêng in Honan, the hydromechanical clockwork dates back about 900 years. (The illustration and others in this article are reprinted with permission from *Science and Civilization in China*, Cambridge University Press.)

clockwork: John Christiansen

Joseph Needham. "Originally we thought of one slim volume, but it just grew like a benign tumor."

completely overturned the usual preconceptions about Chinese thought and civilization and has lifted Joseph Needham and his collaborators to a pinnacle of scholarly achievement. It is a work that one critic calls "perhaps the greatest single act of historical synthesis and intercultural communication attempted by one man." Published a volume at a time by Cambridge University Press beginning in 1954, *Science and Civilization in China* now stands at nine fat books (two more in press) and more than 6,000 pages. Scholars from around the world have contributed, but most of those 6,000 pages were written by Needham himself. "Originally we thought of one slim volume," Needham says wryly, forty-five years later, "but it just grew like a benign tumor."

To start, Needham and his colleagues drew up a master plan "dividing the universe into seven volumes"—a notion that would have delighted Diderot and the rest of the eighteenth-century French encyclopedists. That is still the plan, but when complete the project will comprise some twenty books, since some of the volumes must be published in as many as six parts. The project is in excellent, seaworthy shape, says Needham at age 81, and it will eventually reach port "whether the original pilots are on the bridge or not."

East to west

Joseph Needham is a tall, large-framed man with silver hair and bright blue eyes. He mounts the stairs between floors at his library with the vigor of a man 30 years his junior. And his voice, as he limns the features of the vast terrain he set out to explore so many years ago, is clear and firm. He begins by debunking.

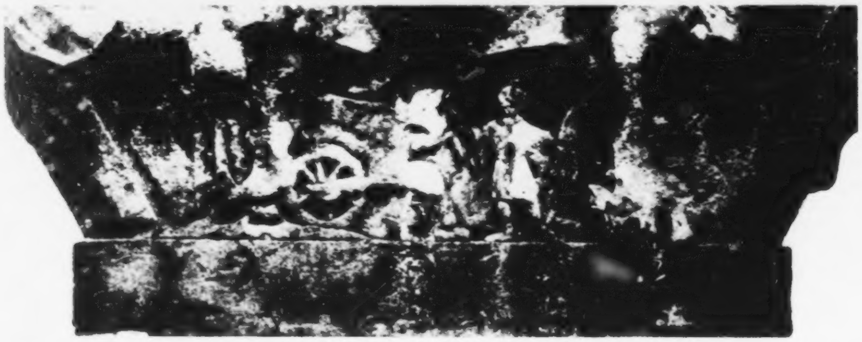
The record, he says, should credit the Chinese with a plethora of scientific and technological feats they are not supposed to have accomplished. "A whole series of bulky volumes" could be written about them, Needham affirms.

These achievements, most of which eventually reached Western Europe either by sea or via the great overland silk caravan routes, include not only the magnetic compass, gunpowder, and printing, but such things as the observation of sunspot cycles, multiple masts and fore-and-aft sails for sailing ships, axially mounted rudders and watertight ship compartments, and a host of mechanical inventions.

Nowhere, Needham declares, is the Chinese record of innovation clearer than in the mechanical realm. He cites, for example:

- The seismograph, invented by Chang Hêng in about 130 C.E. and used for centuries thereafter.
- The simple wheelbarrow, invented in the third century C.E., but not to appear in Europe until a thousand years later.
- The equatorial mounting and the clock drive for astronomical observations. A famous astronomical clock tower, constructed by So Sung at Kaifêng in 1088 C.E., has armillary spheres and celestial globes that were rotated by water power, using a constant-level tank, a driving wheel with buckets, and a kind of escapement or ratchet. It was thus an early example of what Needham calls hydro-mechanical clockwork.
- The first great segmental arch bridge, built in China by Li Chun soon after 600 C.E. There were none in Europe until after 1300 C.E., in Italy.
- The first suspension bridge, using iron chains, built in China at least as early as the sixth century C.E. The principle was not suggested in Europe until the end of the sixteenth century, and the first European bridge of this type was not built there until the late 1700s.
- Foot or boot stirrups did not appear in the West until the eighth century C.E. but are seen in Chinese tomb figures of about 500 years earlier.
- Both principal kinds of efficient harness for draught animals—the collar and the postilion. They appeared in China centuries before they did in the West.
- The essential components of the modern steam engine. These were seen in China at least as early as the sixth century C.E., with the combination of crank and connecting rod for the interconversion of rotary and longitudinal motion. The double-acting principle, used in the reciprocating steam engine, cannot be traced in Europe back beyond 1700. But Chinese double-acting piston box bellows are described in a book published about 1270, and it is probable that they were in use more than a thousand years earlier.

In some cases, Needham says, there was no connection between a western invention and a much earlier similar or identical one in China. This would be true of Fraunhofer's reinvention of the clock drive in 1824. But in general, he adds, "technical inventions...show a slow but massive infiltration from east to west throughout the



Wheelbarrow. The wheelbarrow was used to transport army supplies in China by the third century C.E. But the use of single-wheeled carriers can be traced even further back to around 150 C.E., as shown in this tomb-shrine relief.

Fisher is group editor for science and engineering at Popular Science.

first 14 centuries of the Christian era."

The actual diffusion of some of these advances from east to west has been well documented. Take paper and printing, for example. Pieces of paper have been found in tombs in the Gobi Desert dating back to 300 B.C.E., and Needham says there is "a veritable timetable" for the route they traveled. "You can pinpoint the papermaking—it's in Baghdad by about 750 or so, and then it comes to Iran and to Egypt and North Africa, and then to Spain. The dates go continuously all along."

Often, however, only subtle clues point to evidence for linkage. The magnetic compass is one example. Its origins probably lie with first-century Chinese priests who used magnetized lodestone for divining boards. "It was clearly known in China and used for navigation," adds Needham, "two or three centuries before the Europeans got hold of it." The European compass could have had Chinese origin, he says, even though Chinese compass needles always pointed south. Although European navigational compass needles pointed north, he notes, there is some evidence that in Europe in the sixteenth and seventeenth centuries, mining and surveying compasses pointed south. That suggests to Needham that "the compass might have traveled to Europe not by sea, around India, but right across the mainland, on the Old Silk Road."

Why not China?

Similar examples of a centuries-long dominance by the Chinese in a diversity of other technologies and sciences—astronomy, chemistry, metallurgy, pyrotechnics—have been developed in massive detail by Joseph Needham and his coworkers, Lu Gwei-Djen, Wang Ling, and Ho Ping-Yü. But to return to the two questions that launched such herculean labors, why was this so? And being so, why did China fail to experience the kind of scientific revolution that changed the face of western civilization forever?

The answers seem to lie in Chinese culture and society, Needham and his cohorts are finding—principally in the failure of a merchant class to appear in China. Mercantilism in China, Needham notes, was stifled by the very system—centralized, agrarian, bureaucratic—that had initially nurtured the rise of science and technology.

To understand this apparent contradiction, says Needham, "one has to realize that the nature of feudalism in China was fundamentally different from that in the West. Instead of being military-aristocratic and territorial, it was bureaucratic. And although at first glance the bureaucratic type of feudalism looks to be much weaker than the military-aristocratic, in fact it was more logical, rational, and much more powerful. And while it seems to have favored natural knowledge in the beginning, it later on completely stopped the appearance of modern science in China."

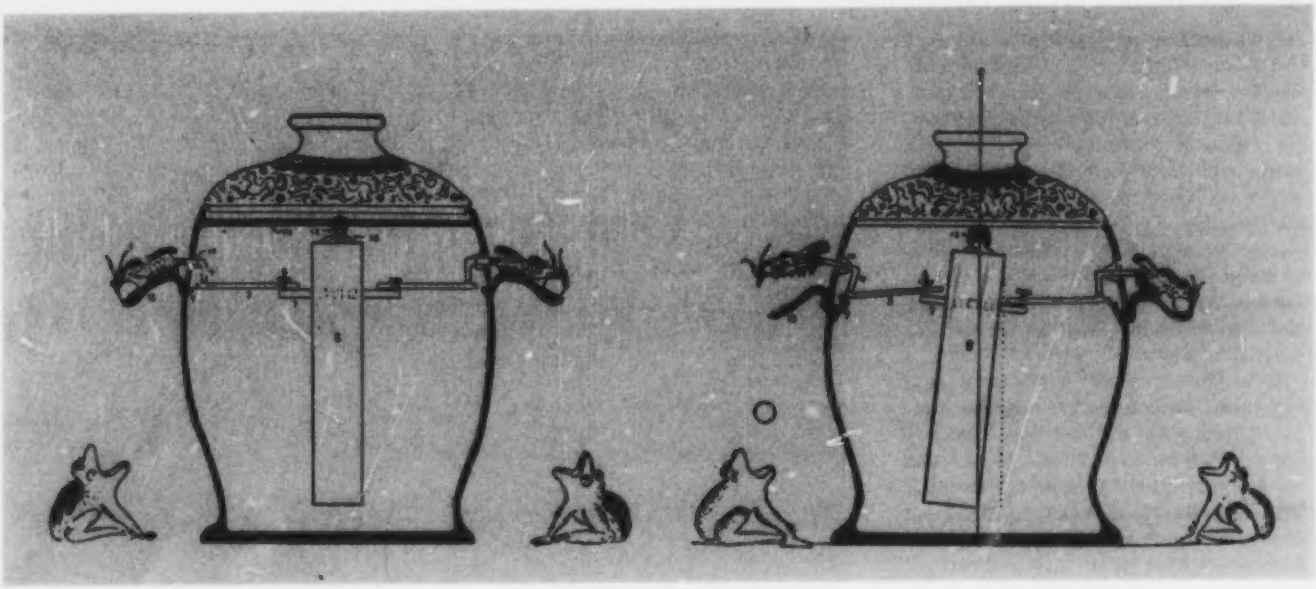
	Invention	Approximate lag in centuries
(a)	Square-pallet chain-pump	15
(b)	Edge-runner mill	13
	Edge-runner mill with application of water-power	9
(c)	Metallurgical blowing-engines, water-power	11
(d)	Rotary fan and rotary winnowing machine	14
(e)	Piston-bellows	*c.* 14
(f)	Draw-loom	4
(g)	Silk-handling machinery (a form of flyer for laying thread evenly on reels appears in the +11th century, and water-power is applied to spinning mills in the +14th)	3-13
(h)	Wheelbarrow	9-10
(i)	Sailing-carriage	11
(j)	Wagon-mill	12
(k)	Efficient harness for draught-animals: Breast-strap (postilion)	8
	Collar	6
(l)	Cross-bow (as an individual arm)	13
(m)	Kite	*c.* 12
(n)	Helicopter top (spun by cord)	14
	Zoetrope (moved by ascending hot-air current)	*c.* 10
(o)	Deep drilling	11
(p)	Cast iron	10-12
(q)	'Cardan' suspension	8-9
(r)	Segmental arch bridge	7
(s)	Iron-chain suspension-bridge	10-13
(t)	Canal lock-gates	7-17
(u)	Nautical construction principles	>10
(v)	Stern-post rudder	*c.* 4
(w)	Gunpowder	5-6
	Gunpowder used as a war technique	4
(x)	Magnetic compass (lodestone spoon)	11
	Magnetic compass with needle	4
	Magnetic compass used for navigation	2
(y)	Paper	10
	Printing (block)	6
	Printing (movable type)	4
	Printing (metal movable type)	1
(z)	Porcelain	11-13

Forerunners. The Chinese developed a range of inventions long before their appearance in the West.

Very early in China's history, Needham explains, the idea that everything in the feudal state should be run by a professional bureaucrat took hold. For 2,000 years, the military was kept under firm control. "It's an amazing thing," he says, "but civilian values were absolutely dominant in China." Every city in China was a node in the administrative network held for the emperor by the civilian governor, Needham explains. There were also military assistant-governors, but their place in the bureaucracy "was about ten ranks down."

This centralized political system, Needham believes, stemmed much less from political ideology than from circumstances, geography being an overriding factor. Chinese civilization began along the Yellow River Valley, and that mighty waterway tends to silt up and overflow, with floods a constant danger. Direct government support from very early times was directed toward river control and irrigation, the making of dikes and canals. To paraphrase Needham, the need for hydraulic works favored the unification of the empire, achieved in 221 B.C.E. And this powerful geographic influence, he says, "equally remains one of the best explanations of why China was traditionally bureaucratic, not aristocratic. Water control and management always tended to transcend the boundaries of feudal domains, since only an emperor and his hierarchy of officials were competent to master it."

This feudal, bureaucratic state was ideally suited to the pursuit of knowledge, especially knowledge that could help improve the human condition. All sorts of laboratories



Seismograph. Reconstruction of a mechanism invented by Chang Hêng in about 130 C.E. records an earth tremor by dropping a ball into the mouth of a frog.

were supported by the state at various times. By the 6th century C.E., there was already an alchemical laboratory kept up by a national academy and by public funds. In fact, says Needham, in the Han dynasty, as early as 60 B.C.E., there had been concerted efforts to produce immortality drugs in the imperial workshops.

Support also went to road builders and bridge designers, with their concern for communications, transport, and the like. And the Chinese state maintained through the ages a national astronomical observatory at Yang-ch'êng in Honan. Like its later western counterparts, it reported on the state of the heavens, on which so much was thought to depend.

Structure, not credo

Why, then, with science and technology flourishing in this relatively benign, powerful, centralized system, did a scientific revolution not take place, either then or later? Or to put it in Needham's own words: "Why did modern science, the mathematization of hypotheses about nature, with all its implications for advanced technology, take its meteoric rise only in the West, at the time of Galileo?"

It is perhaps worthwhile first to dismiss the elements, some superficially credible, that appear decidedly not to have been responsible for the failure of such development in China. Religion, for example was not a factor. Religion in China, Needham points out, was actually much less important than it had been in India. Needham describes Confucianism, the ascendant creed, as "very this-worldly." And Taoism, its principal competitor, "didn't recognize anything outside nature; the Tao was the order of nature, the way nature works."

In the Tao (literally, the way) "there wasn't anything at all inimical to science," says Needham. "In fact the Taoists were always around when scientific discoveries or technological innovations were made in ancient China." As for Buddhism, the most unworldly of the Chinese religions, "[it] couldn't really have set itself against science at all, because in a way it wasn't interested." Indeed, he adds, "some of the greatest scientists in China in the Middle Ages were Buddhists—I-Hsing, for example, the greatest mathematician and astronomer of his age."

Weakness in mathematics? Not a factor. The Chinese had invented decimal place-value, with a blank space for zero, long before anyone else—sometime in the fourth century B.C.E. "In the thirteenth century," says Needham, "they were far more advanced than Europe in the solution of numerical equations. They never got to calculus, but I think their mathematics was quite good enough to have allowed a scientific revolution."

Nor were they restrained by any philosophical credo from performing experiments, an essential ingredient of scientific progress. Needham points to "that great triumph of the ceramics industry in China," the invention of porcelain. "All those glazes they made, and the infinite number of their artistic achievements, couldn't have been done without an enormous amount of experimentation and record keeping about how to arrange the kiln, how to produce oxidizing or reducing atmospheres (as we would call them), what glaze to use, and so on. The same is true of silk weaving and pattern weaving. I think in a way the magnetic compass is also a case in point, because the earliest one was a lodestone spoon. And that was derived from a kind of proto-chess, where you threw pieces representing the planets and other things onto a board and saw where they fell—divination by throwing. So experimentation wasn't a missing element."

What was missing, in Needham's view, was the mathematization of hypotheses about nature and the theoretical constructs it made possible. Mathematization, combined with unremitting experimentation, was the keystone of the scientific revolution in the West. It is what might be called the Galilean method. Despite the presence of mathematics, "discovery of the best method of discovery" just did not take place in China, says Needham.

A package deal

Galileo, born in the year of Shakespeare's birth and Michelangelo's death (1564), personifies the intellectual leap we call the scientific revolution. And mathematization was perhaps his greatest contribution.

"Philosophy is written in that great book which ever lies before our gaze—I mean the universe—but we cannot understand it if we do not learn the language and grasp the symbols in which it is written." In this famous passage from the *Assayer*, Galileo continues: "The book is written in the mathematical language...without the help of which it is impossible to conceive a single word of it, and without which one wanders in vain through a dark labyrinth." It was the essence of the Galilean method to construct precise theories about natural phenomena—the motion of bodies, for example—and then interrogate them, as it were, with conscious, experimental tests.

"Two thousand years before," science historian Alexandre Koyré has written, "Pythagoras had claimed that number is the actual essence of things, and the Bible later taught that God had based the world upon number, weight, and measure. Everybody repeated this, but nobody believed it. At least, up to Galileo nobody took it seriously."

Certainly not the theory-poor Chinese. Needham characterizes Chinese science and technology as "Vincean" because of the very theoretical backwardness of that astonishing figure Leonardo da Vinci. For example, da Vinci believed, without ever testing the notion, that the moisture of a wet rag had an intrinsic tendency to move toward fire. "It helps us to understand," according to Needham, "what remarkable technical achievements may be effected without adequate scientific theory."

But what was it that happened in Renaissance Europe that never happened in China? Why did Europe mathematicize the universe and the Chinese not?

"It's very clear that in Europe there was a sort of package deal," Needham explains. "The scientific revolution took place along with the rise of capitalism and the Reformation. They may be somewhat difficult to disentangle, but none of them took place in China at all." And capitalism, the rise of a mercantile class, seems to be at the crux of the matter. This was also the view of a leading philosopher of capitalism, the late Joseph A. Schumpeter, who held that the emergence of business calculation in the late Middle Ages led to the development of scientific rationality. "In this respect," he wrote, "it is highly significant that modern mathematical experimental science developed, in the 15th, 16th, and 17th centuries, not only along with the social process called the rise of capitalism, but also outside the fortress of scholastic thought and in the face of its contemptuous hostility."

Dead end

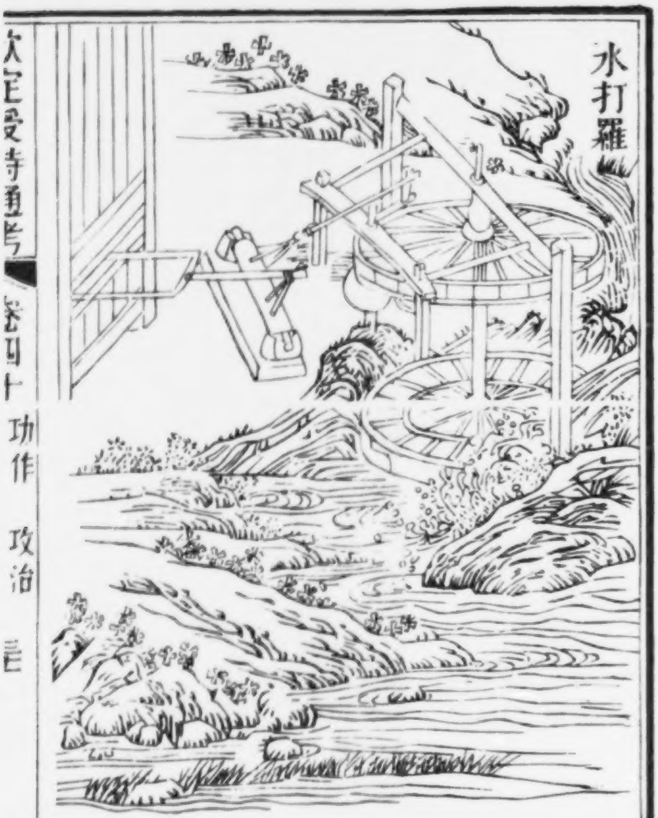
But how would the needs of a merchant class impel a Galilean approach? Primarily because it needed the substitution of a world of quantity for a world of quality. Mercantilists absolutely had to quantify, says Needham. "If you bought a hundred barrels of some particular kind of oil in some island of the Aegean, then you absolutely had to know its properties, the weight of it, the spreading power, what it would dissolve."

Of course, the Chinese also had merchants. "But they were not in close relation to the state power," says Needham. "They were always tremendously kept down in China. They could acquire great wealth, but they couldn't become gentlemen. For that you had to know the classics and get into the civil service. The disparity led to a general despising of all things that interested merchants, and therefore of quantitative science. It isn't really easy to unravel the exact relationship.

"But in Europe, where buying and selling were extremely important, and the people who bought and sold were in influential places in the state, you got a great interest in methods of measurements of all sorts. And that links up with mathematization, or asking exact questions of nature."

To amplify his point, Needham compares technical developments that should have paralleled each other in China and the West. The Chinese should have been interested in mechanics for ships, hydrostatics for their enormous network of canals, and ballistics for guns (having possessed gunpowder four or five centuries before Europeans). "If they were not," Needham asks, "couldn't the answer be sought in the fact that little or no private profit was to be raised from any of these things in Chinese society, dominated by its imperial bureaucracy?"

In the absence of the mercantilist need to quantify everything from astronomy to horse-



Basic mechanics. The hydraulic sifting or bolting machine (left) is based on the same principles as the metallurgical blowing engines, using the combination of crank and connecting rods for the interconversion of rotary and longitudinal motion.

Ancient bridge. The oldest segmental arch bridge in the world was built in China around 600 C.E.

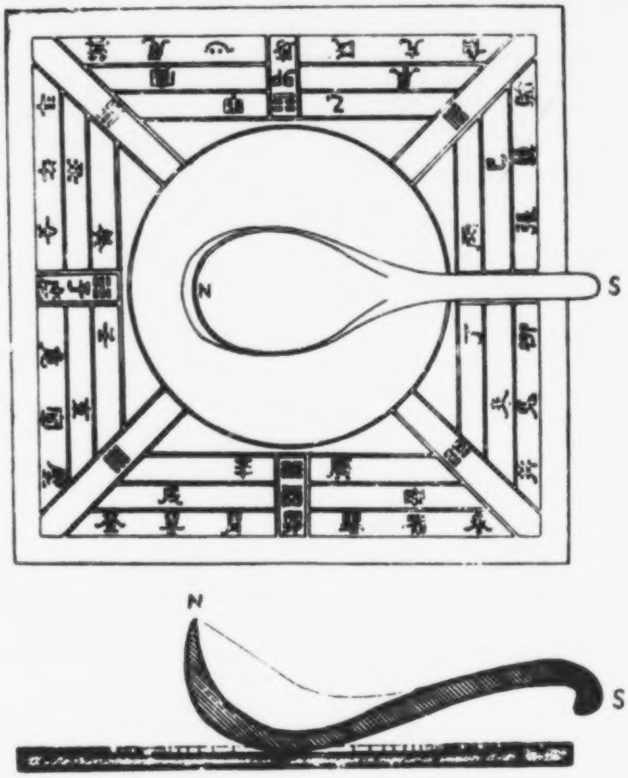
Mao I-Sheng

power, and in the absence of mercantile ascendance, Chinese science hit a dead end, Needham summarizes: "Interest in nature was not enough. Controlled experimentation was not enough. Empirical induction was not enough. Eclipse prediction and calendar calculation were not enough. All these the Chinese had. Apparently a mercantile culture alone was able to do what agrarian bureaucratic civilization could not: bring to fusion point the formerly separated disciplines of mathematics and nature knowledge."

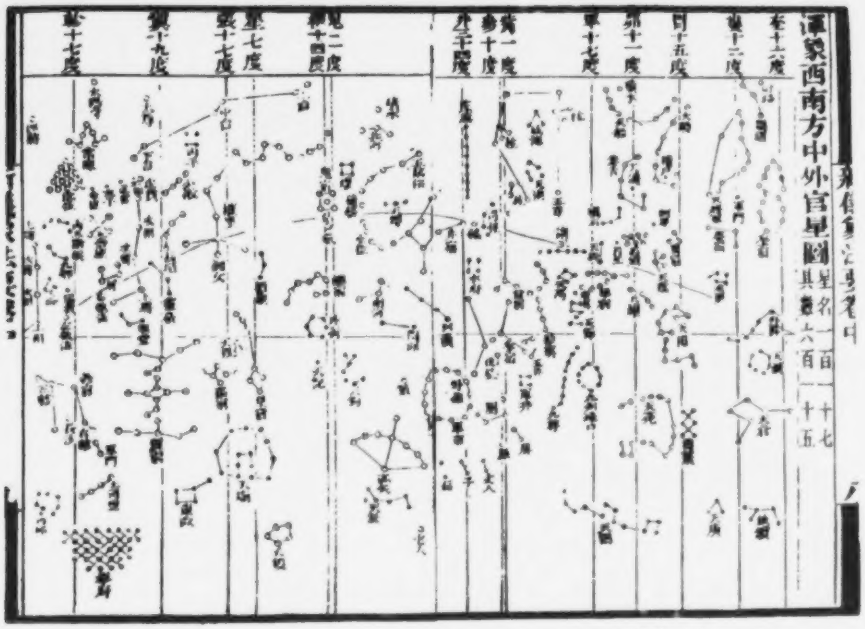
Disparate views

Not all scholars—not even all of Needham's present collaborators—embrace this viewpoint wholeheartedly. An example is Nathan Sivin of the University of Pennsylvania, a specialist in several areas of early Chinese science, who contributed much to the general theory of elixir alchemy. (The search for an elixir of life, described in the thirteenth century by Roger Bacon, was a notion that Europeans acquired from China.)

Of differences between his and Needham's perspectives, Sivin says, "The main difference is that Needham is concerned with finding out what particular discoveries, looked at in isolation, deserve credit from the standpoint of modern science." But since Sivin started much later, when the issue of credit had already been resolved, he has been concentrating on how the Chinese themselves reasoned and how they worked out their procedures. As Sivin puts it in his contribution to the latest of the *Science and*



Lodestone spoon. The compass might have developed out of lodestone spoons such as these, used on a divining board.



Star map. The Mercator projection of the sky was incorporated about 1092 C.E.

The project

Alfred North Whitehead once observed, in *Science and the Modern World*, that "Chinese civilization is the most voluminous in the world." To deal with that incredibly diverse civilization, Joseph Needham has assembled a host of contributors and collaborators from the United States, England, China, Japan, Sweden, Australia, Canada, Hong Kong, France, Germany, and elsewhere. Most are specialists in a single area of Chinese science or culture as well as general sinologists. All are continuing their own scholarly work while they participate in Needham's joint venture: the multivolume *Science and Civilization in China*. Contributors from the United States include:

- Derk Bodde, Emeritus Professor of Chinese at the University of Pennsylvania, writing on the world outlook of the traditional Chinese scholars.
- Ch'ien Ts'un-Hsuin, Emeritus Professor and Regenstein Librarian of the University of Chicago, on paper and printing.
- Peter Golas, Professor of Chinese at the University of Denver, on mining.
- Huang Jen-Yu, Professor of East Asian History at the State University of New York at New Paltz, on fiscal, economic, and social history.
- The late Lo Jung-Pang, Emeritus Professor of History at the University of California at Davis, on military technology, the salt industry, and deep borehole drilling.
- Joseph McDermott of Harvard University, on demography.
- Nathan Sivin, Professor of Chinese at the University of Pennsylvania, on theories of Chinese alchemy.
- Robin Yates of the Chinese Library at the University of Chicago, on military technology.

The headquarters for this international effort is the East Asian History of Science Library, a modest, three-story house near the Botanic Gardens in Cambridge, England. It houses, besides books, a unique collection: the thousands of pieces of research material, in a profusion of languages, that are essential to the project. Here Joseph Needham works and writes ten hours a day, together with a rotating complement of collaborators and assistants that includes Lu Gwei-Djen, his chief collaborator. Support for the project comes from many sources (United States and British aid is from the National Science Foundation and the Cambridge University Press respectively); it is channeled through the East Asian History of Science Trust in Cambridge, Chicago's East Asian History of Science, Inc., and a parallel organization in Hong Kong. •

Civilization in China volumes: "I was interested in something he really wasn't, to seek out what the Chinese alchemists themselves thought they were doing. So there is a kind of complementarity in our interests."

Sivin believes that it is distracting at the least, and probably futile, to expect any answer to the questions about the missing scientific revolution in ancient China "until the Chinese tradition has been comprehended from the inside." He holds, despite Needham's doubts, that an internalist or ideological explanation will have to be found before Needham's monumental effort will really explain the failure of modern natural science to arise in China.

This is too idealistic for Needham. "One must always expect," he says, "that some of these intellectual limiting factors will be definable. But for my part I remain skeptical that there are many factors of this kind which could not have been overcome if the social and economic conditions had been favorable for the development of modern science in China."

In the long run, it may be that the most valuable part of the immense contribution made by Needham and his associates has been to establish for China (and by inference for all cultures) the grand ecumenical nature of science and technology. "My collaborators and I," says the Cambridge scholar, "have long been accustomed to the image of the ancient and medieval sciences of all peoples and cultures as rivers flowing into the ocean of modern science. But there is room for a great deal of difference of opinion on how the process has happened, and how it will proceed. ...It will be a long haul before we can answer the questions of the Sphinx about the development of the natural sciences...and it will take the combined efforts of all of us to do it."

It also seems likely to Needham that some future toiler on that long haul will wander down an obscure byway already explored somewhere in the millions of words of *Science and Civilization in China*. Needham recalls one such encounter of his own: "It reminded me of the triumphal wooden gateway built by Wang, the Taoist priest. He put it upon the path through the barren mountains from the Thousand Buddha Caves to the temple of the Goddess of Mercy by [illegible] running springs. In that desolate, [illegible] place in Kansu there is the inscription '*Ni ye liao ma*': What, have you come here too?" •

The National Science Foundation supports its part of the research discussed in this article principally through its History and Philosophy of Science Program in the Division of Social and Economic Science.

REDEFINING THE COSMOS

by Henry T. Simmons

Assaulting the ageless question of the nature of the universe, cosmologists have advanced a view they call anthropic. The notion bends the line where science and philosophy meet.

Several aspects of our universe—some of which seem to be prerequisites for the evolution of any form of life—depend rather delicately on apparent "coincidences" among the physical constants.

B. J. Carr
M. J. Rees
Nature
April 12, 1979

...what we can expect to observe must be restricted by the conditions necessary for our presence as an observer.

Brandon Carter
Cracow
1973

...we should have learned by now that ideas that appear at first sight to be vague and mystical sometimes turn out to be true.

Freeman Dyson
Princeton
1979

Around 1510, Nicolaus Copernicus began to circulate drafts of his theory of a heliocentric solar system. The ultimate result: a revolution in the scientific probe into natural phenomena. The Copernican cosmology, in which the earth was but one of half a dozen or so planets orbiting the sun, was vastly simpler and more satisfying than the cumbersome geocentric cosmology of Claudius Ptolemy.

Even more important than the stripped-down celestial mechanics of the Copernican model was its tacit assumption that the earth and its inhabitants are not singled out for special attention in the universe. In the absence of hard evidence to the contrary, Copernicus was saying, the earth should be regarded as a typical rather than unique piece of the firmament. This principle of mediocrity, as it was called, was dismaying in the light of contemporary theology and the anthropocentric notions of the Renaissance. Still, it provided the vital impulse for scientific investigation of the natural world. And that inquiry was suddenly unencumbered by a presumption that the world must conform to a particular divine plan, to Aristotle's description of it, or to some other construct.

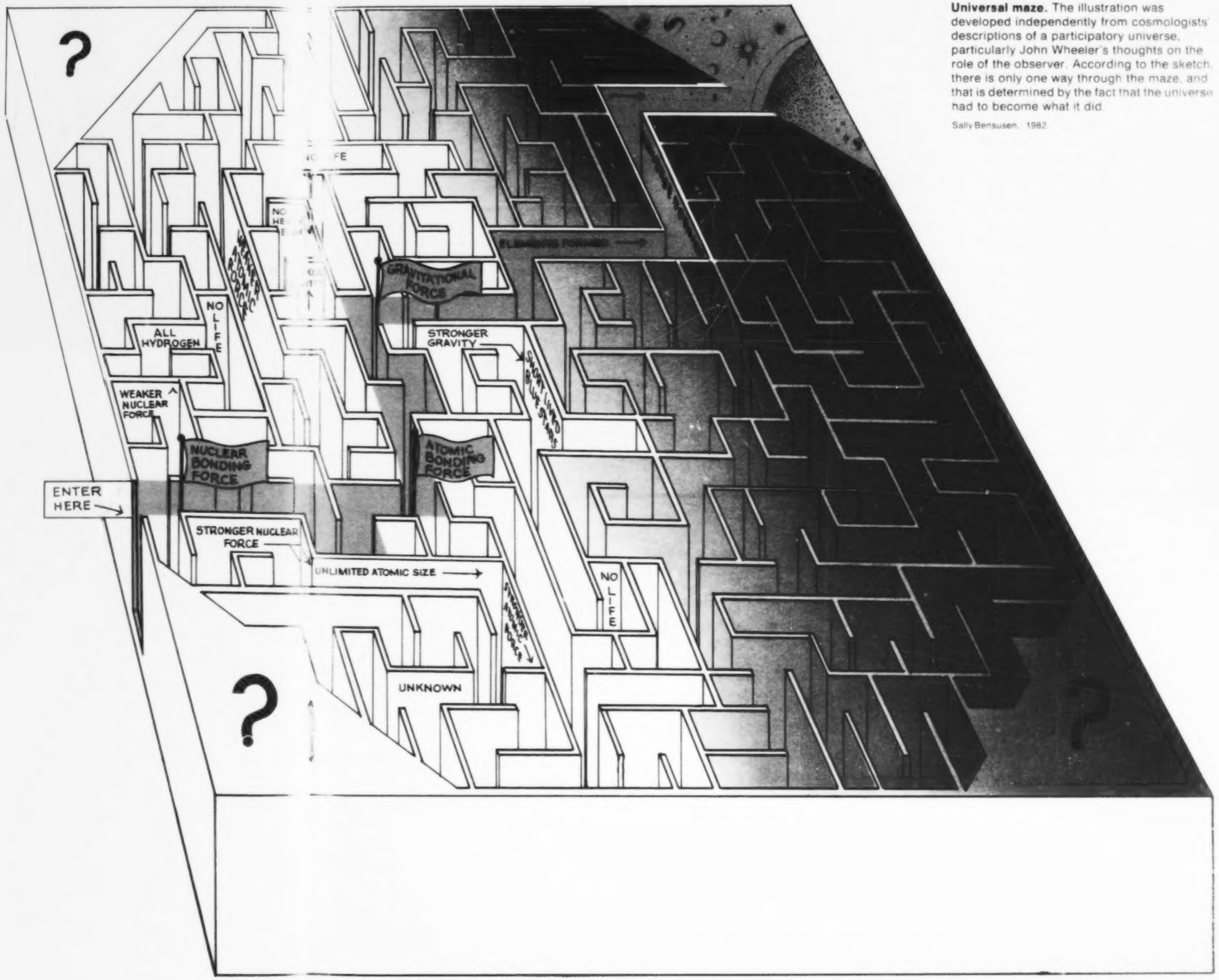
More than four centuries after Copernicus, human culture benefits from a truly prodigious amount of knowledge about the universe and its processes—the fruit of untrammeled scientific investigation. However, there are also prodigious gaps, contradictions, and what seems to be an impressive array of coincidences in the way the universe and its interlocking forces are seen to hang together.

Made restive by some of these coincidences, a small band of American and British astronomers, cosmologists, and physicists has begun an assault. They are challenging what one of them—Brandon Carter, now of the Paris Observatory in Meudon, France—has called "exaggerated subservience to the Copernican principle." They find themselves nudged into a view of the universe that has come to be called anthropic; they see a post-Copernican universe whose nature is constrained by the fact that it contains observers.

Coincidences

Consider just a few coincidental happenstances:

- The coupling strength of the gravitational force, one of those dimensionless numbers physicists use to describe the universe, is so delicately poised that a slight change would alter radically the nature of the cosmos. To some scholars this shift would rule out the possibility of intelligent



Universal maze. The illustration was developed independently from cosmologists' descriptions of a participatory universe, particularly John Wheeler's thoughts on the role of the observer. According to the sketch, there is only one way through the maze, and that is determined by the fact that the universe had to become what it did.

Sally Bensusen, 1982

observers. If the gravitational force were slightly greater, all stars would be blue giants. Such stars manufacture the elements heavier than iron essential to all life on earth and scatter them into space in supernova explosions. But the lifetime of blue giants is only a few tens of millions of years, probably too short for planetary life to evolve and far too short for the evolution of intelligent observers. (In the earth's case, at least, considerably more than 3.7 billion years elapsed between the appearance of life and the emergence of intelligence.)

But if the gravitational force were slightly smaller, all stars would be hydrogen-burning dwarfs, like the sun. Their lifetimes would be many tens of billions of years. This allows ample time for any evolution that might occur. But such a universe would be devoid of the essential-to-life elements heavier than iron.

Thus the value of the gravitational coupling seems precisely poised to permit the evolution of a particular universe. This universe must contain short-lived, metal-scattering, blue-giant stars; long-lived, evenly burning, slowly turning stars such as the sun; and observers.

- There is a similar delicate balance in the strong force that binds the nuclei of atoms together against the enormous repulsion of their like-charged protons. If the strong force were slightly weaker, multiproton nuclei could not hold together and only hydrogen could exist in the universe; if slightly stronger, nuclei of almost unlimited size might exist. Similarly small changes in the value of the electromagnetic coupling constant or in the ratio of the mass of the electron to that of the proton would block any conceivable kind of chemistry.

The weak force also plays a critical role. "If [the weak-interaction coupling constant] were slightly smaller or larger...helium production would be either 100 percent or zero," observe Bernard Carr and Martin Rees of Cambridge University. In the one case, there would be no water, in the other an entirely variant stellar evolution.

- Additionally, the expansion rate observed in the universe today is very close to the critical value necessary to achieve an open, flat, or infinitely expanding universe. The universe, as the cosmologists say, is very nearly if not exactly flat, and Princeton physicist Robert Dicke notes that this flatness is crucial for its evolution. "If the rate of expansion in the early universe were only one part in 10^{14} smaller," he observes, "the universe would have recollapsed before it would have formed stars and galaxies. And if this expansion rate were increased very slightly, by only one part in 10^{14}, the universe would expand too rapidly to permit density fluctuations in the early universe to condense into bound systems like galaxies."

Steven Hawking and Barry Collins, two theoretical physicists at Cambridge University, stated their case in a 1973 paper: "It is only in those universes which have very nearly the escape velocity that one could expect galaxies to develop, and we have found that such universes will in general approach isotropy. Since it would seem that the existence of galaxies is a necessary condition for the development of intelligent life, the answer to the question 'why is the universe isotropic?' is 'because we are here.'"

It was Brandon Carter who in 1973 threw down the gauntlet on behalf of a universe constrained by nature to hold observers. In a paper delivered in Cracow on the 500th anniversary of the birth of Copernicus, Carter declared: "Copernicus taught us the very sound lesson that we must not assume gratuitously that we occupy a privileged, central position in the universe. Unfortunately, there has been a strong tendency to extend this to a most questionable dogma: that our situation cannot be privileged in any sense. This dogma...is clearly untenable...if one accepts that specially favourable conditions are prerequisite for our existence, and that the universe evolves and is by no means spatially homogenous on a local scale....

"I am now convinced," Carter declared, "...that...these coincidences should be considered as confirming conventional physics and cosmology [and] could in principle have been used to predict them all in advance of their observation. However, these predictions do require the use of what may be termed the anthropic principle, to the effect that what we can expect to observe must be restricted by the conditions necessary for our presence as observer." Although the human situation is not necessarily central, it is inevitably privileged to some extent, Carter argued.

Related constants

In proposing a universe influenced in its nature by the fact that it is observed, Carter was building on an earlier proposition by Robert Dicke of Princeton University. Dicke was probably the first to introduce the anthropic principle into physics and cosmology when he proposed in 1961 an alternative explanation of what is called the large-numbers hypothesis. The hypothesis had been advanced by the British mathematician P.A.M. Dirac to account for some strange coincidences among a group of implausibly small and implausibly large dimensionless numbers: the gravitational fine-structure constant (10^{-40}), the age of the universe (10^{40}), and its mass (10^{80}). (See "The riddle of large numbers," accompanying this article, for the derivation of such values.)

To Dirac, the relationship of the gravitational coupling interaction to the age of the universe should be quite independent of that age. Dicke argued that the two numbers must relate to one another. In an evolutionary universe originating in a Big Bang event, Dicke declared, the universe's age "is not permitted to take one of an enormous range of values. [It] is somewhat limited by the biological requirements to be met during the epoch of man."

For example, the universe must be at least old enough for organized galaxies to form and for stars to manufacture, distribute, and reincorporate heavy elements into second-generation stars. As Dicke puts it, "It is well known that carbon is required to make physicists." On the other hand, if observer-physicists are to exist, they must have a "hospitable home in the form of a planet circling a luminous star." The duration of such main-sequence stars is a few tens of billions of years.

But where Dirac's large numbers come in, the important thing, Dicke argued, is that between the minimum and the maximum time observers can exist, the coincidental relationships cited by Dirac must apply. Thus, declared Dicke, contrary to Dirac's original supposition, the age of the universe "is not a 'random choice' from a wide range of possible choices, but is limited by the criteria required for the existence of physicists."

The two important numbers here are the age of the universe and its square, the number of particles within the limits of what can be seen in the universe (equivalent to its mass as measured in protons). Explains Princeton astrophysicist Richard Gott, "If you estimate the number of protons within this big sphere, you notice it is the square of the age of the universe. But you also notice that as the universe grows older, the visible horizon or radius of the universe expands, you can see more of it, and there are more and more particles coming into view."

In other words, the fact that the number of particles in the universe seems to be approximately the square of the universe's age is a simple consequence. It comes out of the fact that the radius of the visible universe is expanding at the velocity of light in all directions and is thus bringing more and more mass into view. The lineup of these two numbers with the gravitational coupling constant just happens to obtain during the epoch of human awareness; it would not

---

Simmons is former science editor for Newsweek.

Participants. John Wheeler, Richard Gott, and Robert Dicke. Assembling an anthropic cosmology.

Princeton University, University of Texas

obtain at very much earlier or later epochs, when human intelligence could not exist. "Thus," says Gott, "the coincidence is explained in terms of people here to observe it."

He offers this illustration: "Suppose the physical constants of the universe were different—for example, that the gravitational coupling were not 10^{-40} but 10^{-20}. Such a universe would have only 10^{40} particles. You might have a universe with 100 stars in each galaxy and 100 galaxies within the horizon. But if there could be observers in such a universe, they would note the very same correspondence that we do in our own more packed universe. I think that this is a very plausible scenario....

"Dirac reasoned that his set of coincidences had to be true at all times," Gott goes on, "but Carter and others say no, that it only occurs now, when intelligent observers are present, just one main-sequence lifetime away from the Big Bang. Dirac required a variation of gravitational coupling with time, but Carter assumes the physical constants are all fixed and that we just see them in this interesting epoch because of who we are."

A coincident density

To Gott, one of the most striking cosmological coincidences is to be found in the density parameter of the universe, called omega. This is another dimensionless number. It relates the actual density of the universe to the critical density needed to reverse its present expansion and thus close it, cause it to cease expanding and recollapse. If the ratio of these two numbers is less than or equal to one, the amount of mass will be insufficient to halt the expansion by means of gravitational deceleration, and the universe will be open. If omega is found to equal one, the universe will have just enough escape velocity to continue expanding forever. In this last situation, the geometry of space-time is said to be flat.

"We calculate a value of 0.06 for omega, and there are other estimates of a value of 0.1 to 0.2," Gott says. "If the universe is open, as these numbers imply, what happens to this density parameter during the evolution of the universe? Obviously it must be very high in the early phase of the universe, say 99 percent, but it must decline asymptotically as the universe expands, reaching zero in an infinite time.

"Now if the universe is indeed open, why are we seeing it at an epoch so close to omega equals one? Why not ten orders of magnitude below the critical density? There is an anthropic reason for this. Galaxy formation tends to occur early in the universe when the density is high, but the formation of new structures like galaxies cuts off when the density parameter falls significantly below unity. The reason is that the density must be sufficient to slow down regions with a density excess of perhaps 1 percent so they can consolidate into bound structures while the universe as a whole continues to expand."

Galaxy formation is no longer occurring today because omega has fallen so far below unity, Gott adds. To have the possibility of condensing into structure, a regional density excess of at least 10 percent would be required today. "The interesting epoch for galaxy formation occurred...when the universe was a tenth of its present size, but subsequent to [the time] when atoms formed and the universe became transparent to radiation. Hence it is entirely logical today that we find omega has only a tenth of the critical value because today's universe is ten times larger than the universe at [the time of galaxy formation]. We should not be surprised to measure omega within one magnitude of the critical density."

A participatory universe

In his Cracow discussion of the anthropic principle, Brandon Carter outlined two versions of it. The weak version, that of Dicke, Gott, *et al.*, merely argues that the conditions prerequisite to human existence sharply constrain the range of possible observable universes. The strong version of the principle goes quite a bit further, arguing that the fine-structure constants and other fundamental properties of the universe are actually constrained to incorporate the evolution of intelligent observers. As Carter puts it, the strong anthropic principle states "that the universe must be such as to admit the creation of observers within it at some stage." Paraphrasing Descartes, he observes: "*Cogito ergo mundus talis est*" ("I think, therefore such is the world.")

Some theoretical physicists, notably John A. Wheeler of the University of Texas, place even greater emphasis on the observer's role in selecting the kind of universe to be observed. Wheeler has long been fascinated with the participatory role of the observer in determining the outcome of what are called delayed-choice experiments. These are experiments in quantum physics in which the observer can decide arbitrarily, while the experiment is in progress, whether the wavelike or particlelike properties of a photon, electron, or some other quantum entity will be observed.

Bridging the distance between the universes of the very large and the very small, Wheeler is thus led to the proposition that the observer and the macrouniverse are also deeply interdependent. The universe, incorporating participatory observers, he concludes, should be regarded as a self-excited circuit. This is a view most other physicists find rather extreme.

In a 1979 paper prepared for the Einstein Centennial at Princeton, New Jersey, Wheeler offered this distillation of the participatory universe: "Beginning with the Big Bang, the universe expands and cools. After eons of dynamic development, it gives rise to observership. Acts of observer-participance—via the mechanism of the delayed-choice experiment—in turn give tangible reality to the universe not only now but back to the beginning. To speak of the universe as a self-excited circuit is to imply once more a participatory universe." Other possible universes, lacking participatory observers, would in Wheeler's view be "stillborn."

For his part, Wheeler has focused on the delayed-choice phenomenon that may be explored in double-slit experiments. He has proposed a variety of arrangements allowing the experimenter to decide what he will observe while a photon or electron is in midflight, but after it has passed through either a single or a double-slit aperture, and therefore after it is presumably committed to exhibit either its particlelike or wavelike property. The outcome of such experiments bears significantly upon notions of time, causality, and a reality independent of the observer.

Says Wheeler:

> After the quantum of energy has already gone through the doubly split screen, a last-instant free choice on our part...gives at will a double-slit interference record or a one-slit beam count. Does this result mean that present choice influences past dynamics, in contravention of every formulation of causality? Or does it mean [us to] calculate pedantically and not ask questions?
>
> Neither: the lesson presents itself rather as this, that the past has no existence except as it is recorded in the present. It has no sense to speak of what the quantum of electromagnetic energy was doing except as it is observed or calculable from what is observed. More generally, we would seem forced to say that no phenomenon is a phenomenon until, by observation, or some proper combination of theory and observation, it is an observed phenomenon. The universe does not 'exist out there,' independent of all acts of observation. Instead, it is in some strange sense a participatory universe.

Other universes

The notion of other possible universes also arises out of interpretation of experiments in quantum mechanics. This notion declares that when a wave function of a particle is collapsed by an observation, only one of the possibilities contained in it is realized; all the rest simply vanish.

Hugh Everett, Neill Graham, and Wheeler have proposed an alternative: that all the possible outcomes become real, but that only one of them appears in our particular branch of the universe. This is known as the many-worlds interpretation of quantum mechanics; in it, the universe branches into a number of parallel universes each time an observation is made.

The Everett picture is entirely consistent with conventional quantum mechanics; it merely bestows on it a more philosophically satisfying interpretation." So Carr and Rees observed in their 1979 review article in *Nature* on the anthropic principle. "There may already be room for the anthropic principle in this picture," they add. "Wheeler envisages an infinite ensemble of universes, all with different coupling constants and so on. Most are 'stillborn,' in the sense that the prevailing physical laws do not allow anything interesting to happen in them; only those which start off with the right constants can ever become 'aware of themselves.'"

Macro and micro

Physicist Freeman Dyson of the Institute of Advanced Studies at Princeton is warmly supportive of Wheeler's bold forays into the unknowable. In a paper delivered to the same Einstein Centennial that heard Wheeler's paper, he declared:

> The idea of observer-participancy is for Wheeler central to the understanding of nature. Observer-participancy means that the universe must have built into it from the beginning the potentiality for containing observers. Without observers, there is no existence. The activity of observers in the remote future is foreshadowed in the remote past and guides the development of the universe throughout its history. The laws of physics evolve from initial chaos into the rigid structure of quantum mechanics because observers require a rigid structure for their operations. All this sounds to a contemporary physicist vague and mystical. But we should have learned by now that ideas that appear at first sight to be vague and mystical sometimes turn out to be true.
>
> Wheeler is building on the work of Bob Dicke and Brandon Carter, who were the first to point out that the laws of physics and cosmology are constrained by the requirement that the universe should provide a home for theoretical physicists. Brandon Carter has shown that the existence of a long-lived star such as the sun, giving steady warmth to allow the slow evolution of life and intelligence, is only possible if the numerical constants of physics have values lying in a restricted range. Carter calls the requirement that the universe be capable of breeding physicists the "anthropic principle." Dicke and Carter have used the anthropic principle to set quantitative limits to the structure of the universe. Wheeler carries their idea much further, conjecturing that the laws of nature are not only quantitatively constrained but qualitatively molded by the existence of observers.
>
> Wheeler unified two streams of thought that had before been separate. On the one hand, in the domain of astronomy and cosmology, the anthropic principle of Dicke and Carter constrains the structure of the universe. On the other hand, in the domain of atomic physics, the laws of quantum mechanics take explicitly into account the fact that atomic systems cannot be described independently of the experimental apparatus by which they are observed. In atomic physics, the reaction of the process of observation upon the object observed is an essential part of the description of the object....[I]t is not possible in quantum mechanics to give an objective meaning to the state of a particle independently of the manner in which the state is measured. Wheeler has made an interpolation over the enormous gap between the domains of cosmology and atomic physics. He conjectures that the requirement of observability will ultimately be sufficient to determine the laws completely. I think he is probably right, though it may take a little while for particle physicists and astronomers and mathematicians to fill in the details in his grand picture of the cosmos.

Hesitant and dubious

It is only fair to say that the great majority of physicists and cosmologists are hesitant and a little dubious about the significance of the anthropic approach, and some are frankly hostile. The weak verson of the principle as outlined by Carter is not the subject of much resistance, but neither does it draw warm support from physicists and cosmologists, who prefer to regard fine-structure constants, mass ratios, and other dimensionless numbers as givens and proceed from there.

"**All of these** dimensionless numbers seem necessary for the universe to evolve as we know it, but there is a circularity in this sort of reasoning and it would be premature to try to attach anything physical to the coincidences," observes astrophysicist David Schramm of the University of Chicago. "Even if we eventually develop a supergrand unified theory of the universe, which incorporates gravity and describes how everything works," he says, "we will still be left with all these 'why' questions, like 'why is this a supergrand unified theory?' Physics tries to answer the 'how' questions, and in some sense it is a philosophical rather than physical undertaking to have a go at these 'why' questions, since they are unanswerable by the techniques of physics."

"**When the anthropic** principle assumes that intelligent life, where it exists, must be similar to human intelligence, it is basically applying a principle of mediocrity," notes Frank Tipler, a theoretical physicist at Tulane University. "It is in this way that the Copernican cosmological principle—that the earth's history is typical—is correctly applied to intelligent life. It is incorrect to use the Copernican cosmological principle to require the inevitable evolution of intelligence on an earthlike planet....To require intelligence on even a large fraction of earthlike planets is basically an attempt to reintroduce teleology—

final causes—into the biology of a single planet." Thus, the weak version of the anthropic principle as formulated by Carter and applied by Dicke would be acceptable precisely because it conforms to the Copernican scientific approach. But Tipler takes issue with the strong version of the principle, on the ground that it imputes a design or purpose in the organization of the universe, one calculated to force the evolution of life and intelligence.

"**The anthropic principle** is an interesting new insight, but it really would take some convincing to rule out the possibility of some other kind of intelligence coming along in an all-hydrogen universe," comments Princeton astrophysicist James Gunn. "The sobering thing to me is the thought that there might be intelligent beings, living in some entirely different physical universe, who might conclude that we were impossible. Hawking and Carter are bright people, and I think there is something important in the idea, but one must be very careful. The basic problem with the anthropic approach is that the arguments are all *posteriori*. But cosmology is not a deductive science, and the arguments are not persuasive."

Gunn is also dubious about the teleological drift of the anthropic case. "A lot of cosmology verges on the teleological," he observes. "When you try to apply physical law to the universe, what is missing is reproducibility, and that is the essence of science. You can have satisfactory explanations in cosmology but not testable explanations. Just where the boundary lies between cosmology as a science and as philosophy is hard to say."

A bridge to carbon

Although it is true that the anthropic approach is essentially retrodictive and *post hoc*, it is generally credited with one predictive triumph. This gives even some skeptics pause, since ability to predict is an important test of scientific theory.

The triumph occurred in the early 1950s, before scientists fully understood how the elements of the periodic table are manufactured by nucleosynthetic processes in stars. The original Big Bang theory, advanced by George Gamow, Ralph Alpher, and Robert Herman in 1948, postulated that all the chemical elements were manufactured in the original fireball. But there was a major obstacle: There are no stable nuclei with atomic masses of five or eight; the ladder is missing rungs.

It was easy to see that protons—atomic mass: one—could capture neutrons—also one—and form deuterons—atomic mass: two—and that collisions of deuterons would produce nuclei of atomic mass four, like those of helium. But the gaps at atomic masses five and eight would have prevented further buildup of elements in the fireball. For that buildup to occur, nuclei of those missing mass numbers had to have been produced, either by the addition of a single neutron to the helium nucleus to form lithium 5, or by the fusion of a pair of helium nuclei into beryllium 8.

It was British cosmologist Fred Hoyle who found a way around this roadblock in 1953. While lithium 5 and beryllium 8 are unstable, they can be produced as brief resonances. Although their lifetimes would be too short in the low-density fireball to enable them to fuse into larger, stable structures before they decay, the conditions obtaining in the dense cores of very hot, helium-rich stars are far different. In the case of the mass-8 barrier, Hoyle showed that collisions between helium

The riddle of large numbers

Thanks to the unified-field theories emerging from high-energy physics in recent years, scientists are beginning to understand why some of the fundamental ratios and constants exhibit the values found experimentally. (See "Grand Unification: An Elusive Grail," *Mosaic*, Volume 10, Number 5.) The theories bring under one tent all the particles and interactions that figure prominently in the quantum domain. By doing so, they explain many of the longstanding mysteries in physics.

Satisfying as these advances are, there are still many fundamental relationships in physics that lie beyond the explanatory reach of unification theories. One of these is the gravitational force, whose effects are negligible in the quantum processes of atoms and nuclei. The coupling constant for the gravitational force is approximately 10^{-40}, astonishingly small compared to all the other physical constants deemed of fundamental significance.

Sir Arthur Eddington was probably the first theoretical physicist to be struck by the incongruity of the gravitational coupling constant. And in 1937 P.A.M. Dirac advanced conjectures based on the remarkable convergence of that constant's vanishingly small value, the age of the universe and the total mass of the universe.

Consider the age of the universe. It is known that the galaxies are receding uniformly in all directions, and at a rate that increases proportionally with distance. Measurements of the red shift of galaxies over a vast distance indicate that recession velocity increases by approximately 20 kilometers per second per million light-years. Figuring backwards from the present, cosmologists can thus calculate that the universe is about 13 billion years old.

To theoreticians, years are meaningless in any fundamental sense. A year represents the period of revolution of but one of nine planets around one of perhaps 10^{22} stars in the universe and thus must be regarded as hopelessly provincial. What is required is a universal way to measure time. That is provided by the classical radius of the electron as a particle. This number is about 10^{-13} centimeters. If the age of the universe is multiplied by the velocity of light in centimeters per second, and that product is then divided by 10^{-13}, another astonishing dimensionless number turns up—about 10^{40}.

This number astounded Dirac. Not only is it far vaster than the other dimensionless numbers in physics, but it also corresponds to the inverse of 10^{-40}, the gravitational constant. Dirac considered an accidental correspondence of these apparently unrelated numbers unlikely. He proposed that 10^{-40} is not truly constant but varies with time; gravity weakens as the universe grows older.

The additional confluence of another dimensionless number was equally staggering for Dirac. If one calculates the mass of the universe in terms of a fundamental unit such as the proton, the resulting number is on the order of 10^{80}. This number is the square of the universe's age and the inverse square of the gravitational constant.

The only way to make sense of these three numbers, Dirac argued, is to suppose that while gravity weakens directly with the passage of time, new mass is created at a rate proportional to the square of the age of the universe and the inverse square of the gravitational coupling constant.

In the years since Dirac advanced his cosmological conjecture, experimentalists have tried to set limits on the rate of change in the gravitational interaction, but the predicted rate is so small that it has eluded definitive test. Meantime, other theorists have advanced an alternative explanation to Dirac's proposal, but one every bit as fascinating: that any cognizable universe must exhibit the relationship observed in Dirac's numbers. This underlies the anthropic principle, in which the presence of intelligent observers powerfully constrains the characteristics of any universe they might observe. •

nuclei are frequent, and the carbon 12 nucleus "wants" so desperately to come into being that a third helium nucleus has a good chance of fusing with beryllium 8 before the unstable beryllium can fracture into a pair of heliums. The result of that threefold union is a 12-baryon nucleus of carbon. And once carbon is formed, the process of helium fusion can go on to build elements all the way up to iron.

Nowhere in Hoyle's paper did he ever conjure the anthropic condition. Nonetheless, his predictive insight is regarded as an anthropic keystone, the bridge to carbon that had to exist if Dicke's observer-physicists were to be.

"Beryllium 8 was the really big stumbling block," recalls Dicke. "You could visualize universes in which the barrier [was not] crossed, for example, by adjusting the values of the strong and electromagnetic coupling constants. From the anthropic point of view, you would still have a universe, but we would not be here. But since we are here, it was obvious that nature had to cross the barrier in some way."

A universe for us

Although Dicke does not press the anthropic principle as hard as Carter or Wheeler do, he favors the idea that there might be an ensemble of universes, a very large collection, as he puts it, and "It is not surprising that we live in one like our own."

Dicke sees it this way:

> Suppose we have just one universe, but one that oscillates. It could be very nearly flat, just barely closed. But in a finite time, all the radiation would be squeezed into a very small volume. The photons would gain energy from gravitation and might produce new nucleons. The next universe would then contain more matter and more entropy—more nucleons, and more photons, but perhaps with the same nucleon-to-photon ratio.
>
> Suppose the universe arose as a few particles created in a quantum fluctuation in the vacuum. After many oscillations, the universe might contain many particles. It might then expand to the size of a walnut and collapse in about a millisecond. But it would bounce, and on each oscillation there would be a new and bigger universe. And somewhere down the line you would finally get to a universe big enough, with a long enough time scale, and we could exist.
>
> But let's just remember the conditions we require. From the fireball we obtain only hydrogen, helium, and a little lithium. Since you can't make people out of only those materials, you have got to arrange for stars to exist, evolve, and scatter the heavier elements out into the universe. But that means a certain minimum time must elapse before the universe makes the elements necessary for life. And you require long-burning stars to give you a little heat.

The elevated observer

In their review in *Nature* of the coincidences supporting the anthropic principle, Carr and Rees marshalled an impressive array of coincidental relationships between the dimensionless numbers and the structure of the universe from the smallest to the largest scales. But sympathetic as they are to the anthropic explanation, they acknowledge its deficiencies: It is almost entirely *post hoc;* it is based on what many regard as an unduly anthropocentric concept of an observer, in that it demands elements heavier than hydrogen and helium, water, galaxies, and special types of stars and planets. "It is conceivable that some form of intelligence could exist without all of these features—thermodynamic disequilibrium is perhaps the only prerequisite that we can demand with real conviction," noted Carr and Rees. Further, the anthropic view does not explain the exact values of the various coupling constants and mass ratios, only their order of magnitudes.

Moreover, the two scientists acknowledged that it may someday be possible to explain, with some as yet unformulated supersymmetry theory that includes gravity, all the things that still appear mysterious and coincidental in relation to their necessity for human existence. "However," they declared, "even if all apparently anthropic coincidences could be explained this way, it would still be remarkable that the relationships dictated by physical theory happened also to be those propitious for life."

And there is another thing. Even if all the coincidences are finally snuffed out by some supergrand unification, what effect if any will the anthropic conjecture have on the human perception of human relationships to the universe? Cosmologists John D. Barrow and Joseph Silk offer this thought: "Whatever the scientific status of the anthropic cosmological principle may be, its impact on the history of ideas may be significant. The principle overcomes the barrier between the observer and the observed. It makes the observer an indispensable part of the macrophysical world." •

The National Science Foundation contributes to the support of the work of scientists whose thought is reflected in this article through several programs in its Division of Physics.

A Pillar of Strength

by Jane Stein

High-strength concrete, used to support high-rise buildings, offers engineers an attractive alternative to structural steel.

The 76-story Water Tower Place in Chicago is dwarfed by the 110-story Sears Tower. And the Helmsley Palace Hotel in New York, at 53 stories, is a pip-squeak among Manhattan skyscrapers. But there is still something special about these two buildings.

They are special because of the way they were constructed: Instead of using structural steel to give the buildings strength, the designers used reinforced concrete, but in a high-strength, slimmed-down version. The largest columns are a svelte meter-and-a-quarter square, less than 75 percent of the bulk of columns made with reinforced concrete of normal strength.

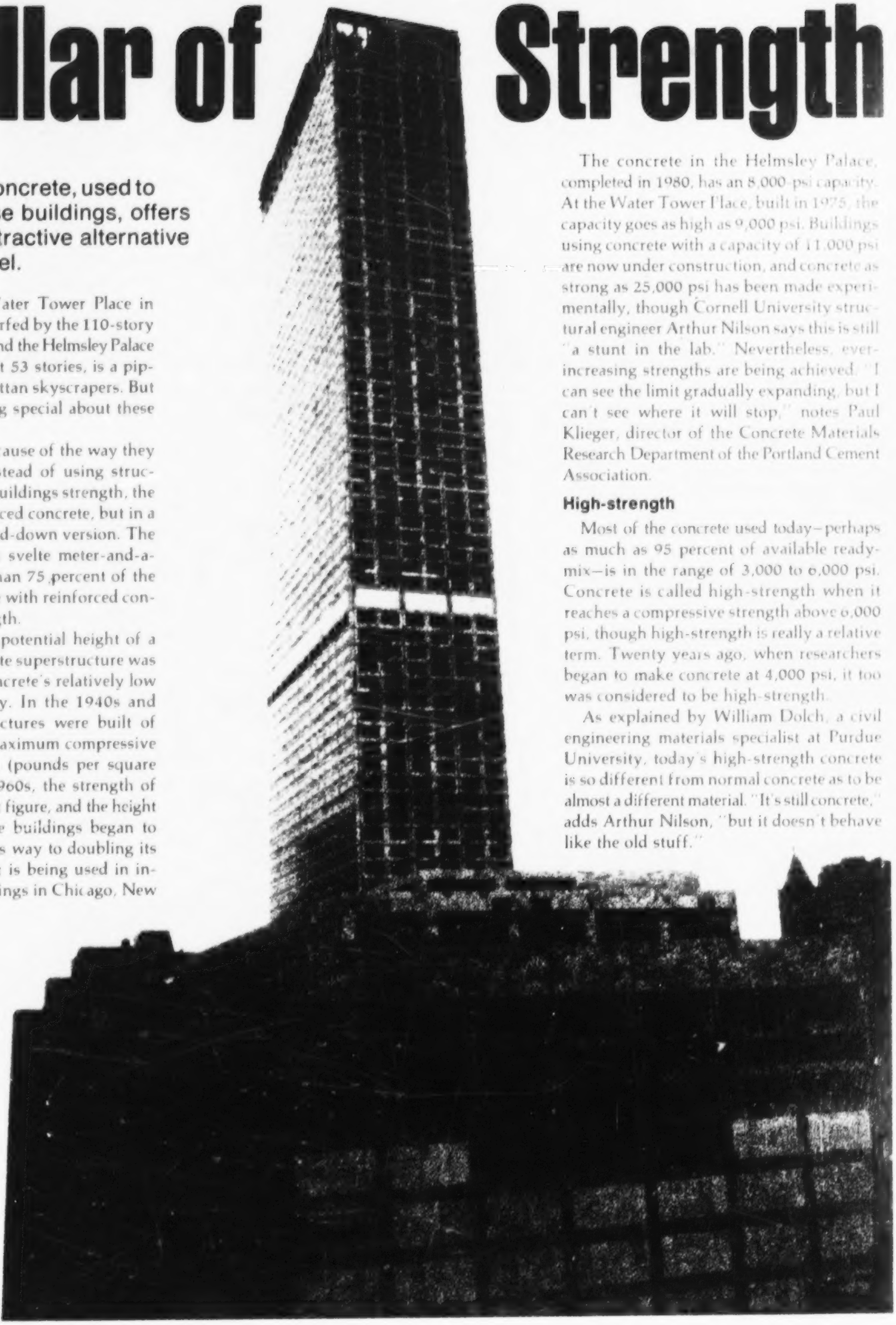
Until recently, the potential height of a building with a concrete superstructure was limited because of concrete's relatively low load-bearing capacity. In the 1940s and 1950s, low-level structures were built of concrete that had a maximum compressive strength of 3,000 psi (pounds per square inch). By the early 1960s, the strength of concrete was twice that figure, and the height of reinforced-concrete buildings began to rise. Concrete is on its way to doubling its strength again, and it is being used in increasingly taller buildings in Chicago, New York, and Toronto.

The concrete in the Helmsley Palace, completed in 1980, has an 8,000-psi capacity. At the Water Tower Place, built in 1975, the capacity goes as high as 9,000 psi. Buildings using concrete with a capacity of 11,000 psi are now under construction, and concrete as strong as 25,000 psi has been made experimentally, though Cornell University structural engineer Arthur Nilson says this is still "a stunt in the lab." Nevertheless, ever-increasing strengths are being achieved. "I can see the limit gradually expanding, but I can't see where it will stop," notes Paul Klieger, director of the Concrete Materials Research Department of the Portland Cement Association.

High-strength

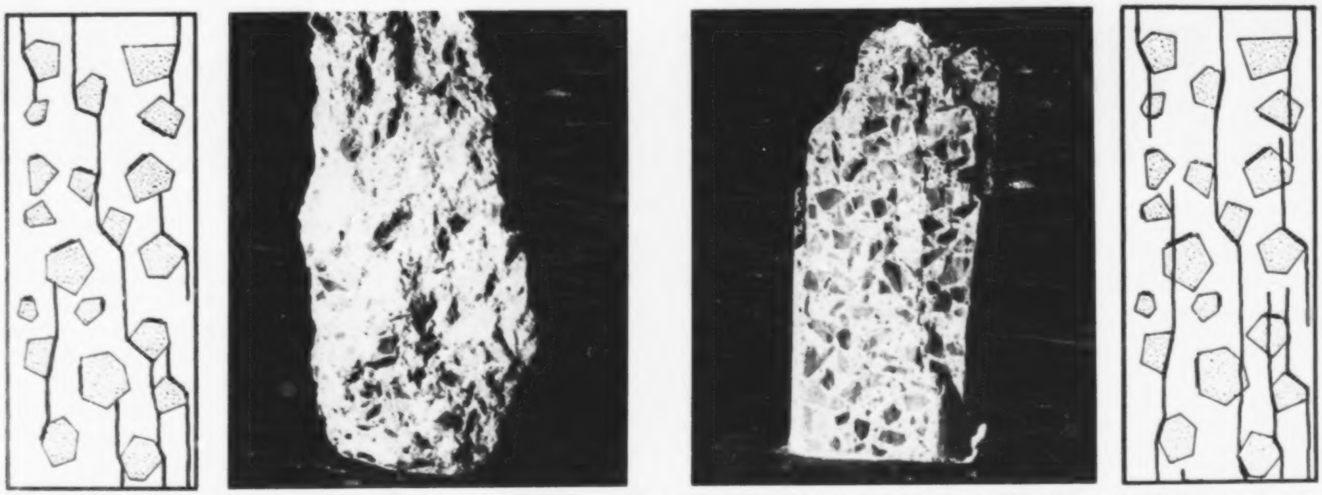
Most of the concrete used today—perhaps as much as 95 percent of available ready-mix—is in the range of 3,000 to 6,000 psi. Concrete is called high-strength when it reaches a compressive strength above 6,000 psi, though high-strength is really a relative term. Twenty years ago, when researchers began to make concrete at 4,000 psi, it too was considered to be high-strength.

As explained by William Dolch, a civil engineering materials specialist at Purdue University, today's high-strength concrete is so different from normal concrete as to be almost a different material. "It's still concrete," adds Arthur Nilson, "but it doesn't behave like the old stuff."



Concrete tower. Water Tower Place in Chicago is the tallest high-strength concrete building in the world.

S. P. Asrow Associates, Ltd.



Normal vs. high-strength. Under compression, normal-strength concrete develops a rough failure surface; cracks pass around aggregates (left). The failure of high-strength concrete is relatively smooth; cracks pass through aggregate and mortar with essentially no change in direction.

S. P. Shah; Arthur Nilson.

Concrete is a seemingly simple material made up of cement paste and aggregates (sand and either gravel or crushed stone). However, it is more complex than it appears, changing its chemical and physical makeup, and thereby its performance, depending on the quantity and quality of the materials used and even how they are mixed.

Source of strength

"Nature didn't make aggregates all alike," Paul Klieger points out. "They differ in strength, elasticity, and surface-bonding characteristics." Some aggregates, for example, are mixed with clay deposits which, if not washed off, weaken the bond between the stones and the cement paste. Shale, a sedimentary rock, can also make concrete weak and vulnerable to high stress. "In high-strength concrete," says Klieger, "you have to recognize what the good sources of aggregates are. With concrete at 3,000 psi it doesn't make as much difference."

Although aggregates can influence the strength of concrete, the ingredient that governs ultimate strength is the cement paste, the glue that binds the aggregate into a rocklike mass. Paste starts out as a mix of water and cement, most of which reacts chemically to produce a gel. In order to mix a workable paste, more than twice as much water as is actually needed for the chemical reaction is usually added to the cement; typically, water contributes half of the weight, though only 20 to 25 percent is needed for the chemical reaction.

The extra water is strictly to make the fresh paste, and thus the concrete, less viscous and easier to work and pour. But it weakens the concrete; when the material hardens, what was excess water becomes pore space. And porous material is always weaker than nonporous.

Excluding water

It has been known for more than half a century that reducing the water content will make concrete stronger. But it was not known how to do this without ending up with a stiff, unworkable mass.

Most of the measures devised to strengthen the load-bearing capacity of concrete are relatively simple and can be performed at the construction site with ready-mix materials. When inexpensive lignin-based compounds—wastes from the paper industry—are added to cement paste and aggregate before or during the mixing process, for instance, water needs are reduced by about 10 percent without any loss of workability. More effective water reducers, popularly called superplasticizers because of the polymer that is added, decrease the water needed by some 25 percent.

There are also mechanical ways to reduce the amount of water needed or to take out excess liquid after the cement has already been mixed. For example, vibrators are used to make drier concrete flow, and vacuum mats suck out water after concrete has been poured.

Changing the composition of aggregates is another way to reduce the amount of water needed. Rounded river gravel uses less water to make concrete workable than do the angular pieces of crushed stone. Having a gradation of rock size, so that small stones fill up the spaces between large stones, can also result in more workability with less water. Taking advantage of a combination of these techniques, the water needed to make high-strength concrete can be brought as low as to 30 percent of the cement paste's weight.

Materials called pozzolans can also be useful in producing higher-strength concretes. Pozzolans are siliceous materials that in themselves have little or no cementing value. But when finely ground pozzolans are added to the cement mix, they react with the calcium hydroxide liberated during cement hydration to form compounds that have strong bonding properties. Pumice and volcanic dust are examples of natural pozzolans. (The word "pozzolan" is derived from products of volcanic deposits from Pozzuoli, in the Gulf of Naples.) Fly ash, a by-product of powdered coal, is a popular artificial product. Pozzolans are either blended into the cement mix at the cement plant or, more frequently, added at the construction site or ready-mix factory.

Factory or field

"There is little need to do any of this for laying a sidewalk," says William Dolch. "But it makes a difference when building a high-rise structure." Because of the detailed attention that must be given to making high-strength concrete, costs are significantly higher than for normal concrete. However, according to the Chicago Committee on High-Rise Buildings, the use of high-strength concrete "is the most economical solution for columns." And columns, they say, are a significant part of the cost of high-rise buildings. The reasons for the economies are that less concrete and reinforcing steel are needed. Additionally, with the high-strength material and smaller-diameter col-

---

Stein, who has also written on biomedical ethics, once worked on the staff of a construction-industry magazine.

umns, there is more space to rent or sell.

"The unique aspect of high-strength concrete is that it can be made that way in the field," says Surendra Shah of Northwestern University. According to the Chicago Committee, concrete at 9,000 psi that was mixed in the field and molded into cylinders compared favorably with similar-strength concrete mixed in a controlled laboratory situation. Although the laboratory-mixed concrete reached the 9,000 figure for strength earlier than the field-mixed concrete did, they both ultimately reached—and surpassed—the desired compressive strength.

There are some operations, however, that require the strengthening work on concrete to be done in the factory. For example, when pozzolans are heated or steam cured, they react chemically with lime, an ingredient that is in the basic cement mix but is left over after the gel is formed. The new compound fills up pore spaces, thus making the concrete stronger, denser, and less permeable. Since it cannot be done at the construction site, steam curing is limited in its actual use to prestressed and precast members.

Other techniques being used in the laboratory to fill in the pores include saturating concrete with hard, super-strength polymers and squashing out pores with mechanical compression, or heating the concrete by compression to attain a pozzolanic reaction.

Low-porosity concrete

Sidney Diamond, in the civil engineering materials department of Purdue University, is working on a low-porosity, high-strength cement that makes concrete in the range of 10,000 to 12,000 psi. Ordinary cement contains a small amount of gypsum to control the setting behavior of concrete. Gypsum, however, causes fine cement particles to clump together, trapping fluid among the clumps. Water then moves with the particles and is no longer free to do its work as a lubricant.

Diamond's problem: to find a substitute for the set-controlling function of gypsum without its flocculating properties. The solution, developed in various laboratories around the world, seems to be to take ordinary cement without gypsum, grind it fine, and add sodium bicarbonate and lignosulfonate. These two ingredients control the set and prevent flocculation. A concrete can be produced that flows into place and is easily compacted even at very low water contents—as little as 26 percent of the cement weight in a mix.

"What we end up with is a tighter structure with much less pore space and a stronger end product," notes Diamond. "Also, it gets its strength at a much earlier stage than does regularly made high-strength concrete. This could be a boon to builders, who will be able to proceed with upper stories faster when building a high-rise." But what Diamond calls low-porosity cement is far from being ready for builders. It is still in the experimental stage; the largest structure built with it to date is a six-meter beam.

Mechanics of concrete

The mechanical properties of normal-strength and high-strength concretes differ significantly. High-strength concrete is reported to be more brittle and less ductile, and it fails in a more precipitous way than does concrete of normal strength. Much of the aim of current research is to determine the limits these characteristics impose on use.

Arthur Nilson, a structural engineer, and Cornell colleague Floyd Slate, a materials scientist, have been interested in the mechanical behavior of high-strength concrete since 1975. They wondered how designers could have the confidence to build structures of high-strength concrete without first knowing how the concrete would behave under a full load. "The rules used in engineering are empirical in many cases, based on experience using concrete in the range of 3,000 to 5,000 psi," says Nilson. "There is no reason to think you could extrapolate these rules to a material that is two to three times as strong."

For example, Nilson explains, the modulus of elasticity, a measure that describes the stiffness of a material, is based on an equation derived from tests on concrete of normal strength. In good-quality concrete, stiffness is controlled by both the cement paste and the aggregate, while the strength is influenced mostly by the paste. In high-strength concrete, one of these components—the cement paste—is substantially changed, and the equation relating stiffness to strength is no longer in balance. If used, the equation could lead to important errors in design, Nilson says.

The Cornell team began to study concrete with experimental compressive strengths ranging from 3,000 to 12,000 psi, using 10-by-20-centimeter compressive test cylinders.

Since concrete is a heterogenous material, its behavior under load can be explained only in terms of the interactions of its different components—specifically the aggregates and the cement paste. These interactions are largely governed by the development of internal microcracks. "Most people think of concrete as solid, but, in fact, cracks spread through it when loads are added," Nilson points out.

Loads and overloads

Nilson and Slate found a fundamental difference in the way regular and high-strength concrete cracked internally. Most of the microcracking in regular concrete is on the interface between the aggregates and the cement. This is because the weakest link in lower-strength concrete is at the aggregate-cement surface. Cracks begin to appear when the concrete is under relatively little load—when it has reached only 30 percent of its full compressive-failure strength. As the load increases, the interface cracks spread slowly and deform the concrete slightly. When interface cracks are connected by cracks passing through the mortar, the concrete finally fails. The fracture plane formed is ragged, following the tortuous routes of the microcracks around the nuggets of aggregate and through the mortar.

When X rays are made of slices of high-strength concrete, they show that there is very little internal cracking prior to failure, because there are virtually no weak links at the stone-cement-paste interface. "In high-strength concrete, the cement is strong, like stone, so that the material is more nearly homogeneous," notes Nilson. When cracks finally develop—and this occurs as far into the test as at 90 percent of compressive-failure strength—they pass cleanly through the stone and paste, rather than following the interfaces. When the material fails, it does so with a sudden, abrupt, and clean crack.

Although there is no warning when this failure will occur, this does not mean that the buildings using high-strength concrete are necessarily dangerous. "Buildings are not loaded to peak," Nilson adds. "They are built with a safety factor of at least two. In addition, reinforcement is provided to reintroduce the ductility that the plain concrete material lacks."

Despite its sudden-failure characteristic, high-strength concrete might be able to carry an overload more safely than normal concrete can. Since it can withstand a higher load without internal cracks developing, when the temporary overload is lightened, the high-strength concrete is intact. With regular concrete, it is less likely that the load can be removed before damage has already occurred, since microcracking begins to develop at a much earlier state of loading.

Explosive or gradual

One of the most important indications of the mechanical behavior of concrete—indeed of any structural material—is its stress-strain curve. This curve relates the stress applied to the compressive deformation that results. A ductile material can withstand the pressure from the full load by gradually deforming before it fails. A brittle material, in contrast, has relatively little deformation after a maximum force is applied and before it snaps.

Researchers agree that high-strength concrete is brittle. But they are still debating when it becomes brittle, and what the stress-strain curve looks like. The descending part of the curve is exceedingly hard to chart because, as Nilson puts it, "Once you are at the top of the curve [the maximum strength], the situation is unstable. The material can deform, carrying a gradually diminishing load, or it can fail even though the load is reduced."

According to Northwestern's Surendra Shah, "When you test a concrete cylinder and it reaches its maximum stress, typically there is a big bang. The stronger the material, the bigger the bang and the more catastrophic and explosive the failure. Engineers want the strength, but they don't want a catastrophe."

This might sound like a death knell for the use of high-strength concrete. But based on tests done while he was at the University of Illinois at Chicago Circle, Shah concludes that the explosive failure is characteristic of the testing machine, not the material.

Shah maintains that the properties of concrete in practice depend not only on its compressive strength but also on the amount of reinforcing steel used. Accordingly, he argues, both materials—concrete and steel—should be tested together. With this testing procedure, Shah hypothesizes, the concrete would not shatter; instead it would lose strength gradually.

To show this, Shah and his research team have team loaded high-strength (11,000-psi) concrete cylinders in parallel with steel tubes. The load was shared simultaneously by both the concrete and steel. Based on experience with 200 test cylinders, Shah concluded that the failure of high-strength concrete is gradual.

Arthur Nilson is not in full agreement with Shah's interpretation. Shah's method, he contends, tests reinforced high-strength concrete, not high-strength concrete itself. "The concrete could be cracked and reduced to crushed stone and would still provide some resistance," Nilson says, but no engineer would want to base his design on concrete in such a state of disintegration. Also, in practice, not all concrete regions are reinforced.

Half the creep

Designers also have to know how high-strength concrete behaves in relation to loads over time. This is particularly important in designing structures such as high-rise buildings and long-span bridges, where most of the load is the weight of the structure itself.

Concrete usually deforms over time when loaded. Deformation over time at a constant load is called creep. At the Cornell labs, Nilson and Slate tested high-strength concrete in cylinders under loads over periods as long as three years. They found that the high-strength concrete deformed a third to half as much as does normal concrete. Nilson has an explanation for this. Creep usually takes place in the cement paste, not the stone, he says. The stronger the paste, the less the creep.

Using normal concrete, a beam of a given size could deflect 2.5 centimeters immediately and 5 more over the course of three years. This is not as serious a structural change as it sounds: Engineers can design a building so that when a heavily loaded column deforms over time, its load is gradually shifted to less-strained columns. Nonetheless, a creep-deformed beam could create functional problems if doors do not close or if sensitive machines do not operate properly because of misalignment.

Creep can also seriously affect prestressing, a factory technique that takes advantage of concrete's compressive strength. In prestressing, concrete beams are formed around taut steel wires. When the ends of the wires are freed, a state of precompression is produced and penalties associated with the concrete's typically low tensile or flexural strength are avoided. With creep, however, tension in the wires diminishes over time because the stressed concrete member is continually shortening.

Knowing that high-strength concrete has much less creep over time could be an important feature in the design of long, prestressed bridges and of wide expanses, such as auditorium roofs. Nilson and Slate are now looking to see what happens to creep on loaded, high-strength concrete beams over a three-year period. "We anticipate that the long-term cylinder tests will correlate with long-term beam tests," says Nilson.

In the field

How well high-strength concrete behaves is revealed in its actual use in buildings. The Portland Cement Association went to the 76-story Water Tower Place, the world's tallest reinforced-concrete building, while it was being built. Cement association researchers, under the direction of structural engineers Henry G. Russell and W. Gene Corley, obtained samples of concrete from ready-mix trucks at the construction site. There were five different strengths of concrete, ranging from 9,000 psi used in the basement to 4,000 psi on the top floors. For each concrete strength, 15-by-30-centimeter cylinders were cast and tested. "We confirmed that the desired strengths were gotten in the field," Corley states.

The association engineers also measured column deformation over time at six different floor levels in Water Tower Place. "Research has shown that the larger the test sample, the less the creep. A cylinder of concrete measuring 30 by 61 centimeters will give a different response than a 15-centimeter-square cylinder of similar compressive strength," notes Corley.

The volume-to-surface ratio appears to be a key to how much moisture can be lost and therefore how much creep there will be. The greater the surface, the greater the water loss and the greater the creep. It follows, then, that tests on the full-sized structural member itself will produce the most accurate measures of deformation in a building. "This is why the studies were done at Water Tower Place," Corley adds. "We also wanted to get answers to serviceability questions [do the ceilings sag or does plaster crack?] and to safety questions [does the column near the elevator shorten so much that it could cause structural damage?]."

In addition to the actual deformation measurements, the association also calculated the degree of shortening based on material properties, stress levels, elasticity, and time of load. In general, Corley points out, there was good agreement between the measured and calculated shortening. By correctly anticipating shortening, designers interested in using high-strength concrete can make appropriate compensations in design at each floor. "A three-millimeter shortening per floor, uncompensated, could make a significant difference in a high-rise structure," says Corley. Tests on Water Tower Place show that the maximum differential movement resulting from column shortening was only 22 millimeters at the thirty-second floor, far less than would affect serviceability or safety appreciably.

The success of Water Tower Place has set people's minds at ease about building skyscrapers with high-strength concrete, notes Sherwin Asrow, a Chicago structural engineer. He predicts that builders will begin using concrete of 12,000 to 14,000 psi on a regular basis. "Research in the laboratories gave us the basis to know what high-strength concrete is, how to get the proper grinds and ideal mixes," Asrow adds.

There are still a lot of research questions about how high-strength concrete works. But builders aren't waiting for all the answers. "A lot of engineering research traditionally goes on in the field, while construction is actually taking place," says Portland Cement's Henry Russell. For high-strength concrete, it will be a combination of laboratory and field work that tells the rest of the story. •

The National Science Foundation contributes to the support of research discussed in this article through its Geotechnical Engineering and Structural Mechanics Programs.

Plasmids: More Than Genetic Debris

by William Check

Ubiquitous rings of genetic material appear to researchers to have a vital role in bacterial virulence and a selfish penchant to replicate.

Among the many elements inside bacterial cells are large, tightly coiled loops of DNA, the cell's genomal helix. When bacteria are broken open and their contents examined under an electron microscope, these large tangled masses of genetic material show up clearly, ultimately to reveal themselves as enormous, tangled rings. But electron microscopy also reveals other, small rings of DNA. These are puzzling because under ordinary circumstances, bacteria can grow quite well without these extra bits of genetic material. Nonetheless, they have been found in almost every bacterial species examined. Eiichi Ohtsubo, of the State University of New York at Stony Brook, calls this situation incredible, and indeed it is hard to imagine that these nearly ubiquitous rings of bacterial genetic material, called plasmids, do not have some important role.

One of the first indications that plasmids do function—if only in their own interest—came in the early 1960s. Japanese microbiologists were studying an outbreak of diarrhea caused by strains of the bacterium *Shigella*, which resisted many common antibiotics. Even worse, the bacteria were able to transfer this resistance easily to other, antibiotic-vulnerable *Shigella* strains. The Japanese microbiologists found that this ominous property was due to plasmids. The

Gene transfer. A bacterium transfers genetic information via conjugation. Plasmids play a key role in this exchange.

Raymond C. Valentine

small DNA loops incorporated the genes conferring antibiotic resistance. Further, they moved easily from cell to cell.

Since then, plasmids have been implicated in many other aspects of bacterial virulence. These include production of diarrhea-causing toxins, the way organisms which are related to plague bacteria escape engulfment by human white blood cells, and the ability of some bacteria to survive other common host defense mechanisms.

Plasmids have also been found to confer survival traits not related to disease on their bacterial hosts—those that can grow in high concentrations of toxic heavy metals and withstand damaging irradiation, for example. Moreover, the plasmid-borne genes that enable *Pseudomonas* bacteria to metabolize hydrocarbons such as xylene, toluene, and octane were the basis for the first patent issued for a bioengineered microorganism: the strain of *Pseudomonas* that Ananda Chakrabarty of the University of Illinois designed to digest oil spills.

Plasmids as plasmids

Plasmids came under close study as molecular entities in the early 1970s, with the advent of restriction-enzyme analysis of DNA and of gene-cloning techniques. They had been studied first for the classic exploratory reason—because they were there—but soon became the vehicle that has carried genetic engineering to its dizzying heights. (See "Recombinant DNA: the Power of the Tool," *Mosaic*, Volume 12, Number 1.) Some investigators took advantage of the restriction enzymes' penchant for cutting DNA at specific sites; they explored the structural basis for similarities and differences of function from plasmid to plasmid. Others cloned particularly interesting plasmid DNA segments to get enough material for code deciphering via nucleotide-sequence analysis.

This work has produced what Richard Novick of the Public Health Research Institute of the City of New York calls "an amazing diversity of discovery about plasmids." Scientists have a new appreciation for the elegant plasmid-host interactions responsible for generating survival attributes for the bacterium and for the plasmids it carries. They have also realized the possible role of the plasmid as a mobile gene pool that may have contributed substantially to bacterial evolution.

In addition, recent research has revealed extraordinary efficiency in plasmid mechanisms for ensuring their own replication and transmission to new hosts. In fact, scientists have begun to think of plasmids as more than hitchhikers bearing one or more genes useful to the bacterial host. The plasmid now looks more like an independent entity that has evolved the capacity to make good use of bacterial replication systems to reinforce its own. In Novick's view, "The plasmid may be the ultimate selfish gene."

Stanley Falkow. "Most plasmids are selfish; they just want to replicate themselves."

William Check

Plasmid ubiquity

Even the plasmid's more obvious role, contributing to a bacterium's ability to thrive under unusual or adverse conditions, was not appreciated for a long time. This was partly a result of the unusual character of the first bacterial plasmid studied: one called F, which resides in *E. coli*. Stanley Falkow, head of Stanford University's medical microbiology department, calls F "a plasmid that has been held captive inside the same bacterial strain in the laboratory since the 1920s and not allowed to do the things plasmids normally do."

In that time, F has developed a seemingly unique ability: It enters the bacterium's main, or genomic, DNA and causes it to transfer that genetic material to another *E. coli* at a high rate—a process that Falkow says he has never seen in a non-laboratory-grown F plasmid. In fact, microbiologists once had concluded that plasmids were rare and that F was a freak. Falkow cites several developments that forced him and other scientists to realize that F is indeed a freak, though only because it can cause transfer of bacterial genomic DNA.

One of the first clues to the ubiquity of plasmids was the emergence of bacterial strains resistant to several antibiotics. These were found in the 1960s in Japan, Europe, and the United States; in all cases, the resistance genes were carried on plasmids.

Meanwhile, other evidence was showing that plasmids were even more widespread

Yagi and Clewell. Yoshihiko Yagi and Donald Clewell have focused their research on plasmids in *Streptococcus faecalis*, a normal resident of the human gut.

University of Michigan

Check is a microbiologist who became a science writer.

than suggested by the instances of antibiotic resistance. In one survey, Falkow collected more than a hundred separate samples of *E. coli* bacteria from humans. He found that approximately 85 percent of them contained plasmids and that most carried three or more of them. Since most of the bacteria were sensitive to antibiotics, Falkow realized that plasmids must code for functions other than antibiotic resistance.

Since then, plasmids have been identified as the virulent element in *Salmonella typhi*, the organism responsible for typhoid, and in scours, a lethal diarrhea in livestock. H. William Smith of Houghton, England, and Carlton Gyles of the University of Guelph in Ontario are the scientists who found that scours was caused by toxins secreted by the bacteria and coded for by a plasmid. They also discovered that the bacteria had a complex mechanism for producing surface proteins. This enabled them to colonize the small bowel, which is normally not heavily populated by bacteria.

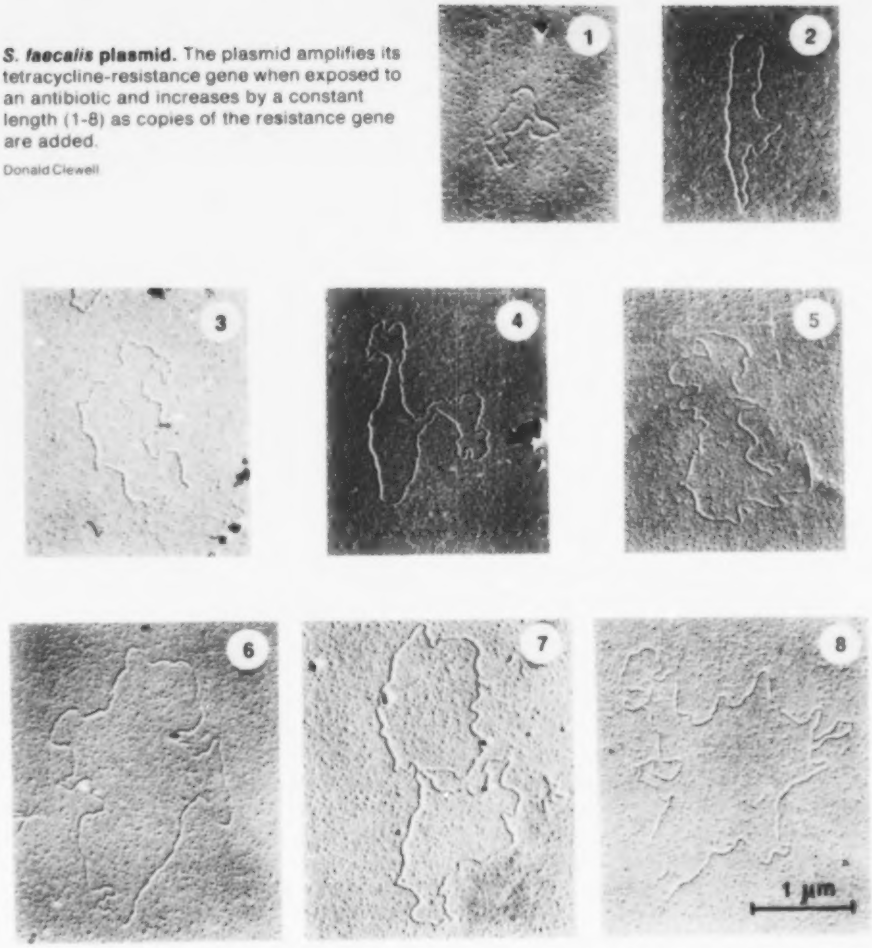
Human diseases

Falkow and his colleagues later identified plasmids with diarrheal disease of human infants and in *E. coli* isolated from individuals with travelers' diarrhea. They also learned that the ability to cause human diarrhea was completly transferable from one *E. coli* strain to another.

Then, in the middle seventies, researchers at the University of Texas and the Walter Reed Army Institute of Research reported that plasmid-carried genes caused the virulence of the bacterium *Yersinia enterocolitica*, a relative of the plague bacillus *Y. pestis*. Dan Portnoy in Falkow's laboratory (then at the University of Washington) found that the plasmid determined the virulence, but the genes in the main bacterial genome directed tissue invasion.

More important, Portnoy found an obligatory relation between these two properties: To cause disease, the plasmid must sense both a particular temperature, 37 degrees centigrade, and the low-calcium ion concentration found inside animal cells. Under these conditions, the plasmid causes the bacterium to synthesize proteins that coat the bacterial surface. This makes the bacterium impervious, as plague organisms are, to phagocytosis by human white blood cells and resistant to the natural killing effect of normal human serum.

Using this information, Falkow constructs a plausible scenario to explain the virulence of *Yersinia* organisms, including *Y. pestis*: The bacteria grow in their intermediate vectors, fleas, without causing too much difficulty. This is because a flea's body temperature is 25 degrees centigrade. When the bacterium invades a human cell, the plasmid responds to the higher temperature of the new host and the low-calcium ion concentration characteristic of the cell's interior. It turns on production of the new surface proteins so that, when the bacterium leaves the human cell, it is equipped to survive the host's cellular defenses.



***S. faecalis* plasmid.** The plasmid amplifies its tetracycline-resistance gene when exposed to an antibiotic and increases by a constant length (1-8) as copies of the resistance gene are added.

Donald Clewell

The temperature sensitivity involved in this intricate cooperation between bacterium and plasmid helps explain another phenomenon: the fever resistance of some disease organisms. Peter Williams, in Leicester, England, and Jorge Crosa, of the University of Oregon School of Medicine, found that some plasmids enable coliform bacteria to grow in the body even when the host attempts, in effect, to starve the invaders with inflammation and fever.

One result of fever and the inflammatory response is the increased binding of free iron in the blood by serum proteins called transferrins. Because bacteria require free iron for growth, this tactic should deprive the invaders. Some of the plasmids—those called Col V, for example—have evolved a counterploy, a small protein that binds iron even better than the transferrins do. The plasmid actually allows the bacteria to snatch the iron back from the transferrins and continue to grow.

Gram positives

Further evidence for the widespread presence and functional diversity of plasmids comes from studies of gram-positive bacteria. All the investigations described so far have concerned gram-negative organisms, which appear red in the gram stain, a venerable experimental reaction devised in 1884 and still widely used. But many medically and economically important bacteria belong to the gram-positive group, which stains purple in this reaction.

This simple red/purple color distinction may have created an artificial dichotomy in the minds of investigators and delayed the discovery of plasmids in gram-positive bacteria. Says Don Clewell, of the University of Michigan: "As recently as 1972, most microbiologists assumed that, except for the streptomycetes, conjugation [a sexual kind of reproduction by which bacteria can transfer genetic material] existed only in gram-negative bacteria." But in the last few years highly specific. So far at least four different pheromones have been identified; each induces production of aggregation substance by bacteria carrying one specific kind of plasmid. When the plasmid enters the new cell, it shuts off the pheromone signal. This system promotes the spread of plasmids while reducing unproductive matings.

Since the pheromones are encoded by the bacterial genomes, Clewell raises the possibility that these peptides had a prior function of possible importance to the bacteria, or even to humans, in whose gut they are naturally found. The plasmids may simply have evolved an ability to take advantage of these preexisting signals.

Analysis of 100 clinical *S. faecalis* isolates in Clewell's laboratory indicated that the pheromone system might be involved in spreading drug-resistant plasmids. Production of pheromones was significantly associated with antibiotic resistance.

Evolution at work

One of the deepest puzzles, likely to stimulate much discussion for years to come, is where plasmids originated. It is difficult to decide even where to look for homologies—in early potential ancestors or later ones. "The Col El plasmid gets by with almost all host functions," Donald Helinski points out. "Does that make Col El primitive or sophisticated?"

The origins of the special genes carried by plasmids are likewise shrouded in mystery. Falkow notes that there is considerable homology between the diarrhea-causing toxins of *E. coli* and those of the cholera organism *Vibrio*. But this does not show which came first, if indeed they are related temporally.

It does seem clear, however, that the existence of antibiotic-resistant plasmids predates the widespread use of antibiotics. When multiple-drug resistance arose in the 1960s, several groups, including Stanley Falkow's, tried to find out whether populations not exposed to the drugs already harbored the plasmids. Each time, in humans in Borneo and the Solomon Islands, in animals in the outback of Australia, the answer turned out to be yes.

Whatever the evolutionary origins of plasmids and their genes, there are several instances in which these systems appear to have undergone important changes in the last 20 years, a mere instant on the evolutionary time scale. Richard Novick has analyzed a group of plasmids from staphylococci that code for antibiotic resistance and heavy-metal detoxification. He found clear sequential relationships among them, with the aboriginal gene apparently being for cadmium resistance and with mercury resistance added early.

At Stony Brook. Eiichi and Hisako Ohtsubo have found the region that directs some RNA synthesis.

William Check

The past five years have provided two examples of how antibiotic resistance can evolve in a previously sensitive species. Since 1975, clinical isolates of the gonococcus responsible for gonorrhea and of *Hemophilus influenzae*, which causes meningitis in young children, have become increasingly resistant to antibiotics. "Both these organisms had been intensively studied for many years," says Novick, "and there had been no evidence of antibiotic-resistant plasmids. It is hard to think that such plasmids would have gone undetected."

Falkow describes two further observations that make him feel "we are watching evolution take place." First, the penicillinase plasmids found in these two bacteria are identical, and in neither case had this plasmid been present before antibiotic resistance arose. Second, the common penicillin-resistant plasmid in *H. influenzae* and in the gonococci lacks a specific DNA recognition sequence found in the DNA of these two organisms, an indication that the plasmid originated in a separate bacterial genus. But what genus? And how did the transfer take place between this putative ancestor and these two very different recipients? These are questions with important implications, but there are no answers yet.

The pre-plasmid

Recent research has also uncovered what may be an evolutionary intermediate gene segment, a conjugative transposon. (See "Gene Segments on the Move," *Mosaic*, Volume 12, Number 1.) Typical transposons are short, tandem DNA sequences that pick up a gene from a genomal DNA and carry it to another site on that DNA or to another piece of DNA. But transposons are ordinarily limited to hopping around between DNA elements within the same cell. To move to another cell they need to recombine—with a plasmid, for instance. Recently, the laboratories of Don Clewell at Michigan and Walter Guild at Duke University have found genetic elements that have the properties of transposons but can induce their own cell-to-cell transfer. They are not plasmids, because they cannot replicate on their own outside the bacterial genome. But this finding raises the possibility that the first step in evolution from transposon to plasmid may have been uncovered: the acquisition of transfer functions, with replication ability coming later.

From an anthropocentric viewpoint, plasmids are both beneficial and detrimental. They enable bacteria to produce useful fermentations as well as to cause disease. For the plasmids, however, there is only one determinant, the need to survive and spread. Bacteria have learned to take advantage of the plasmid's genetic peregrinations to further their own evolution. Scientists have also used bacterial plasmids in their genetic engineering feats. Humans may eventually find that they too have plasmids and perhaps will turn the tables, learning to use them for their own benefit. (See "Plasmids in vertebrates," accompanying this article.)

Whatever course humans take, it looks as though plasmids will probably adapt easily, continue to propagate, and show a few more genetic tricks in the process. No one knows where they came from, but it seems that plasmids are here to stay. •

The National Science Foundation contributes to the support of research discussed in this article principally through its Genetic Biology Program.

Evolution

I have just read Boyce Rensberger's article "The Evolution of Evolution" (*Mosaic*, Volume 12, Number 5), and felt that I must write to express my delight. Rensberger's article provides the best review of the current debates and new ideas in evolutionary theory that I have read. A superb job of reporting and writing. He succeeds nicely in "slicing through the morass and putting everything in perspective." It's so nice to see a careful, rational synthesis of the latest ideas of the molecular biologists, the paleontologists, the developmental biologists, and population geneticists.

Well done!

Alan N. Bunner
Perkin-Elmer

The National Science Foundation & Avery Publishing Group

are proud to present

The Mosaic Reader Series

The genius of the biological scientist

As scientists probe the mysteries within us and our world, **THE MOSAIC READER SERIES** focuses on and highlights current significant research at leading university, industry and government laboratories. This timely and authoritative series will present scientists, educators and students with a concise, intelligent overview of the advances being made by preeminent scientists around the world.

AN AUTHORITATIVE SERIES

The first five books available in this fascinating series focus on research in the specialized fields of the biological sciences. The volumes in this set are:

THE CELL: Inter- & Intra-Relationships
DNA: The Master Molecule
HUMAN EVOLUTION
EVOLUTION: New Perspectives
ECOLOGY: Impacts and Implications

When completed, **THE MOSAIC READER SERIES** will contain more than twenty-four volumes encompassing research in the five major scientific areas: Biology, Anthropology, Physics and Astronomy, Chemistry and Earth Science. This comprehensive reference library of current scientific research will examine topics ranging from the topology of DNA to stellar ontogeny.

Each volume is a collection of articles written and edited by professional science journalists and authenticated by the scientists actually engaged in the research being discussed. These articles, originally appearing individually in *Mosaic* magazine, have been carefully evaluated and specially selected for the breadth and lasting significance of the material they cover. Many of the articles contain [illegible] which highlight topics of special interest that are related to the principal subject. The articles are further enhanced by numerous diagrams, illustrations and photographs. Each volume is indexed and contains a complete glossary.

To order copies of the titles currently available in **THE MOSAIC READER SERIES**, complete the attached order card at the back of this *Mosaic*.

National Science Foundation
Washington, D.C. 20550

Official Business
PENALTY FOR PRIVATE USE, $300

Postage and Fees Paid
National Science Foundation

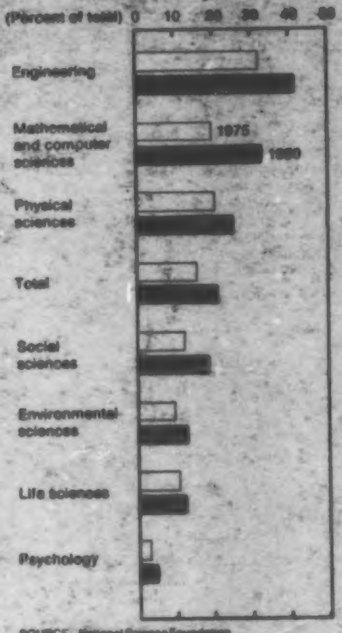
U.S.MAIL

BULK RATE
U.S. POSTAGE PAID
PERMIT NO. G69
WASHINGTON, D.C.

Foreign Students Are One-Fifth of Total Science/Engineering Graduate Enrollment in U.S. Doctorate-Granting Institutions

- Since the mid-1940s, the United States has provided graduate training, especially in the sciences and engineering, to an increasing number of students from overseas. Developing nations face severe problems requiring scientific and technical personnel in far greater numbers than their own higher education systems can produce. Because foreign governments have increased their support of students studying abroad, the number of foreigners enrolled in U.S. institutions of higher education has grown significantly.
- The growth in the number of foreign students was concentrated in the science/engineering fields and at the graduate level.[1] While the number of science/engineering graduate students with U.S. citizenship enrolled at doctorate-granting institutions grew by less than 3 percent over the 1975-80 period, the comparable number of foreign students grew by 47 percent.
- Over the 1975-80 period, the proportion of full-time foreign graduate students in science/engineering increased from 16 percent to 21 percent. The 1980 share was substantially greater in engineering than in the sciences—42 percent compared to 17 percent. The ratio was highest in petroleum engineering (57 percent). Of the full-time graduate students in metallurgical/materials engineering and agricultural engineering, nearly half were foreign. This reflects the high priority assigned to the development and exploitation of natural resources.
- In general, foreign graduate students made up smaller proportions in science than in engineering; there were significant exceptions, however. In statistics, 42 percent of the graduate students were foreign and in economics 38 percent. In psychology, by contrast, foreign students were only 5 percent of the total.
- Although foreign enrollment in graduate science/engineering programs at doctorate-granting institutions continued to grow through the late 1970s, the number of doctorate degrees awarded to foreign students declined, as did the number going to United States citizens.
- In most fields, the proportion of all degrees awarded to foreign students declined. But in engineering, they increased; foreign students received 46 percent of the doctorates in 1980, up from 41 percent in 1975. This is largely a reflection of the declining number of United States citizens staying on to complete their doctorates in the face of the high starting salaries offered by industrial firms to baccalaureate and master's-degree holders. The only other area in which foreign students received a larger share of the doctorates awarded in 1980 than in 1975 was mathematical/computer sciences, in which industrial demand for personnel was also high.
- Foreign participation in American higher education ultimately affects the employment situation both in the United States and in the students' home countries. One measure of the relative strength of these two influences is the proportion of doctorates awarded to foreign students in the United States on permanent visas (immigrants) compared to those on temporary visas (nonimmigrants). In 1975, about one in three doctorates awarded to foreign students went to a student on a permanent visa. By 1980, the ratio had fallen to one in four. The relative decline between 1975 and 1980 occurred in all fields except the social sciences (including psychology) and the mathematical/computer sciences.
- Although nonimmigrants comprise the bulk of foreign students receiving science/engineering doctorates in the United States, they have little effect on the United States labor market in general. Students from foreign countries may have a substantial effect on the curriculum, however, especially in those fields where they make up a significant part of the total enrollment. There is no positive indication yet of any substantial modification of curricula to meet specific needs of foreign students. However, there have been occasional complaints that some course material is inappropriate in terms of third-world concerns.[2]
- Those foreign Ph.D.-holders who plan to remain in this country have a direct effect on the United States labor market, though the small numbers involved tend to minimize this effect in most fields. Foreigners in the United States on permanent visas received only 4 percent of the Ph.D.'s awarded in science fields in 1980. In engineering the proportion was 12 percent—down slightly from the 14 percent reported in 1975 but still a significant share of the total. In view of the shortage of doctorate-level engineers in all sectors of the economy—especially as faculty in academic institutions—recruitment from abroad of engineers trained in the United States may fill a growing number of vacancies on engineering faculties.[3] By 1990, foreign students could make up as much as one-sixth of the doctoral engineering labor force, in the absence of a substantial growth in the number of United States citizens staying in graduate school to complete their doctorates.[4]

Foreign as a percent of total full-time graduate science/engineering enrollment in doctorate-granting institutions by area



SOURCE: National Science Foundation

Full-time graduate enrollment in doctorate-granting institutions and number of Ph.D.'s awarded by citizenship and science/engineering field: 1975 and 1980

	Total	Engineering	Physical[1]	Mathematical[2]	Life[3]	Social[4]
Full-time S/E graduate enrollment:						
Fall 1975	210,307	37,106	30,238	14,182	43,409	85,170
Fall 1980	230,801	42,451	32,386	16,279	44,604	72,864
Foreign:						
Fall 1975	33,128	11,863	5,236	2,818	5,161	6,860
Fall 1980	48,748	17,695	6,901	6,004	6,397	10,087
Ph.D.'s awarded:						
June 1975	18,486	2,861	3,831	1,147	4,540	6,209
June 1980	17,185	2,478	3,151	963	4,710	6,890
Foreign:						
June 1975	4,015	1,223	813	272	868	839
June 1980	3,576	1,149	682	238	777	831
Nonimmigrants:						
June 1975	2,758	810	544	197	596	611
June 1980	2,642	850	506	162	591	584
Immigrants:						
June 1975	1,257	413	269	75	272	228
June 1980	934	299	176	76	186	247

[1] Includes environmental sciences.
[2] Includes computer science.
[3] Excludes medicine.
[4] Includes psychology.

SOURCE: National Science Foundation and National Research Council

---

[1] National Science Foundation, *Foreign Participation in U.S. Science and Engineering Higher Education and Labor Markets* (NSF 81-316) (Washington, D.C.: Government Printing Office, 1981), p. 1.

[2] See Association of American Colleges (AAC), *International Education Trends 2000*, "Educational Policies in the Middle East," Vol. 1, No. 4 (Washington, D.C., 1979), p. 106.

[3] A recent Higher Education Panel report indicates that about 10 percent of all full-time engineering faculty positions were vacant in fall 1980. See Frank J. Atelsek and Irene L. Gomberg, American Council on Education, Higher Education Panel Report No. 52, *Recruitment and Retention of Full-Time Engineering Faculty, Fall 1980* (Washington, D.C., October 1981), table 1.

[4] National Science Foundation, *Foreign Participation*, p. 17.

All issues available have
been filmed.
Unfortunately, any
scattered missing issues
are no longer available.

READEX FICHE best copy available

13/4

NATIONAL SCIENCE FOUNDATION

MOSAIC

VOLUME 13 NUMBER 4 JULY/AUGUST 1982

L[redacted] C
NEWARK PUBLIC LIBRARY
FILE ROOM

Summer storms

Trails in the sea

Superluminal motion

Plant evolution

Integrated optics

Ice in the Sky

NS1.29:

In this *Mosaic*

Unlike articles in issues of *Mosaic* that are built around a common theme, the articles in issues like this one are selected for the commonality they lack. The range is intended to represent the variety of scientific areas with which NSF shares a boundary. Each such general issue can show only a fragment of the spectrum, but as the volumes accumulate, we do manage to cover the ground.

Despite the randomness of the selection, however, occasionally articles come together in non-theme issues that share an idea, often one that is important to science. There may be such a pair here.

The articles come from disciplines as far apart as archeology and astrophysics: Sam Iker's article on the origins of agriculture, "The Seeds of Agriculture," on page 2, and Edward Edelson's on superluminal motion, "Faster Than the Speed of Light?" on page 24. What they have in common is that, in each, a long-held tenet of science is questioned. It is not likely that these assaults on conventional thinking will be equally successful; indeed, neither may succeed. But they do make the points that in scientific research nothing is (or certainly ought to be) beyond question and that questioning is the process by which knowledge is advanced.

In the case of "The Seeds of Agriculture," what is being questioned is the well-entrenched notion that the invention of agriculture in the Levant, no more than 11,000 years ago, was the seminal event in the emergence of civilization. The challenge comes in the form of two assertions: One is that agriculture was evident at least 7,000 years earlier in an area far to the southwest of the Fertile Crescent. The second is that the first was a non-event: It took 7,000 years for the invention to convert itself into a revolution, for people who had learned how to manipulate plants to decide to settle down and do it in one place.

The other assault is even more fundamental. It would dislodge a keystone of virtually all known laws of physics: that the velocity of light is a constant. Astronomers whose instruments have achieved the power to resolve details of quasars and galaxies have seen them flying apart at speeds greater than the speed of light. If what is seen is truly there, the ramifications are limitless.

Needless to say, both sets of observations—the archeological and the astrophysical—are being examined with great and critical care.

From the lay point of view, whatever the outcome—and Iker and Edelson present the pros and cons quite clearly—it ought to be not only fun to watch, but instructive as well. Scientists defending what they have come to accept are often forced to think in new ways, and science can almost be seen to be advancing as they do.

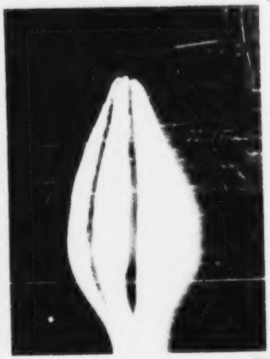
Warren Kornberg

MOSAIC

VOL. 13, NO. 4 JULY/AUGUST 1982

2

10


30

NATIONAL SCIENCE FOUNDATION

Lewis M. Branscomb, *Chairman*
National Science Board

John B. Slaughter, *Director*
National Science Foundation

Thomas Ubois, *Acting Director*
Office of Government and
Public Programs

Warren Kornberg, Editor
David Heffernan,
Assistant to the Editor
Joyce Latham, Copy Editor

Mosaic is published six times yearly as a source of information for the general public and the scientific and educational communities served by the National Science Foundation. The Director of the National Science Foundation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this agency. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1984. Subscriptions are available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription is $12.00 per year in the United States and possessions and $15.00 all other countries. Single copies are $3.25 domestic and $4.10 foreign. Send check or money order to the U.S. Government Printing Office. Communications other than those concerning paid subscriptions should be addressed to: Editor, *Mosaic*, National Science Foundation, Washington, D.C. 20550. Telephone: 202-357-9776.

Cover. A radar dome dominates the silhouette of an operations center for thunderstorms research in Miles City, Montana. (See "Ice in the Summer Sky," page 10.)

Henry Lansford

The Seeds of Agriculture *by Sam Iker* 2
Evidence found at desert sites in Egypt suggests that rudimentary agriculture began there some 18,000 years ago. Some archeologists remain skeptical.

Ice in the Summer Sky *by Henry Lansford* 10
In a massive, coordinated effort, scientists are exploring the dynamics of summer storms. How ice forms in storm clouds and in turn influences cloud formation is high on the list of questions.

A known mechanism 14

Born to Navigate *by William J. Cromie* 17
Many marine species come armed with senses that enable them to perform feats of navigation for which humans need batteries of sophisticated instruments.

Faster Than the Speed of Light? *by Edward Edelson* 24
The appearance of superluminal motion observed in some quasars and galaxies must be an illusion. If not, the laws of physics will have to be rewritten.

To Land, To Seed, To Flower *by Boyce Rensberger* 30
Plants had to conquer the earth's rocky terrain to set the world stage for the animal life that came along next.

Integrated Optics: A Communications Revolution *by Mort La Brecque* 35
Integrated optical devices will do much that integrated electronic circuits now do. The problem of the optical switching transistor may set the limit.

Optical switching transistors 39

National R/D Expenditures Expected to Reach $85 Billion in 1983 Back Cover

The Seeds of Agriculture

Evidence found at desert sites in Egypt suggests that rudimentary agriculture began there some 18,000 years ago. Some archeologists remain skeptical.

by Sam Iker

The letter from Egypt arrived on an early spring day in April of 1978, and archeologist Fred Wendorf's 12-year search came unexpectedly to an end. "I have a very nice surprise for you," declared Cairo University botanist Nabil el Hadidi. "Your samples from Wadi Kubbaniya included four grains of cereal, and I feel sure that they are barley."

The samples sent to Cairo for analysis had been extracted by Wendorf's expedition from the barren sands of the upper Nile Valley. And he had no idea that any of them contained seeds. Indeed, the disappointed archeologist had returned to Southern Methodist University in Dallas several weeks earlier, after fruitlessly combing prehistoric sites near the city of Aswan for just such evidence.

With the startling news from his Egyptian colleague, an elated Wendorf felt convinced that his theory was finally confirmed; he was certain that the nomadic hunters and gatherers who lived some 18,000 years ago on the fringes of Wadi Kubbaniya, an ancient river bed leading into the Nile, were practicing a primitive sort of agriculture.

Earlier, Wendorf had found large numbers of grinding stones and crudely fashioned stone sickles at another upper Nile prehistoric site. These, he believed, were used to harvest and process grain. But he and his colleagues had come up with no traces of the grain itself. Now there was grain, too.

Because of the climate of Wadi Kubbaniya and the physical location of the latest critical find, Wendorf felt sure that the newly discovered barley could not have been growing wild. (The nearest wild stands today are located some 800 kilometers to the northeast.) It meant to him that the prehistoric nomads of Wadi Kubbaniya were planting as well as collecting and grinding the grain in a rudimentary form of farming.

A great nonevent

Wendorf's hypothesis is heretical, at least, in the challenge it poses to existing theories of the origins of agriculture and its importance. If he is right, scientists around the world will have to reexamine their notions of what led to this watershed change in human lifestyle and the place it ought to be assigned in the emergence of human civilization. That is not a challenge to be taken lightly.

It has long been widely accepted among archeologists that the first agriculture began some 8,000 to 11,000 years ago, far to the northeast of Egypt, in the Fertile Crescent that ran through Lebanon, Syria, Iran, and Iraq. It was thought—and still is by many scientists—that proximity to large stands of wild cereal grasses, environmental stresses, or population pressures—or a combination of them—had motivated primitive peoples to shift from a wandering way of life to sedentary farming. Scholars and scientists considered this the watershed in human history that led to urban civilization.

Wendorf found no evidence of any such proximity, environmental stresses, or population pressures at Wadi Kubbaniya. Nor were his farmers sedentary. "If our interpretations of the...evidence are correct," he declares, "then it would seem, in this case at least, that the development of food production was one of the great nonevents of prehistory."

This sort of statement is guaranteed to raise hackles among archeologists, many of whom have spent their entire careers in the Mideast probing the origins of agriculture and its place at the base of subsequent social evolution. Predictably, many questioned Wendorf's interpretations of the evidence. Even after he and his coworkers had unearthed more than 200 additional seeds at Wadi Kubbaniya in 1981 (including lentils, chickpeas, dates, and wheat, as well as barley), many of his colleagues still resist his view of agriculture as a great nonevent.

Grinding stones and sickles

Wendorf first raised the possibility in 1966 that the prehistoric, nomadic occupants of the upper Nile Valley were practicing a primitive form of farming thousands of years before agriculture began in the Levant. He had come to Egypt several years earlier to direct a joint United States-Polish-Egyptian expedition surveying Stone-Age remains that would be flooded by the water rising behind the Aswan Dam. "We were dealing with 400 square miles of reservoir," Wendorf recalls, "and literally thousands of archeological sites."

One of the areas he and his associates excavated and studied in 1964 was located at Tushka, on the west bank of the Nile, near the temple at Abu Simbel. The site is a large wadi encompassing a number of late Paleolithic sites. Wendorf dates them back about 14,000 years. (The Paleolithic, or Old Stone Age, period ran from more than 1.75 million to 10,000-12,000 years ago.) In those sites, Wendorf and his colleagues found a number of grinding stones. "Initially, the significance of these grinding stones didn't occur to me," Wendorf recounts. "We were assuming the people were harvesting occasional wild grasses. But among the stone artifacts we also found a few flake pieces which had sharp edges with a high gloss." Wendorf describes these blades as primitive sickles.

It was that discovery, according to Wendorf, that set him to making connections. "Why do we have sickles from 14,000 years ago? Why are all these grinding stones here? It suddenly occurred to me that there was something more going on." In an exhaustive literature search, he found that such stones had been common in the Near East 10,000 years ago. "But 14,000 years ago there were virtually none."

Up to that point, archeologists, including Wendorf, had felt the grinding stones found at Tushka and elsewhere in the upper Nile Valley would have been used by prehistoric peoples to prepare body paints or to process wild grasses. The idea that they were crude instruments of domestic agriculture didn't occur to them. "You have to look at the mind-set at that time," Wendorf explains. "No one would have thought that Africa or the Nile Valley—this area—could have been an early center of plant domestication." Not

First hint. Archeologists at first resisted the idea that grinding implements uncovered at Wadi Kubbaniya were crude instruments of early agriculture.

Fred Wendorf/*Loaves and Fishes*

only is the area extremely arid, but "all of the wild relatives of the early domestic plants grew far away," Wendorf observes. Further, the inhabitants at the time were known to be nomads, not farmers. "So we looked for all kinds of other explanations."

Why agriculture?

Food production should have suggested itself earlier, says Wendorf. "From back before World War II," he declares, "the question of how, when, and why food production began is one that has occupied anyone who has had any interest in our past and how societies have developed. Every civilization known on earth has had as its fundamental economic core the ability to produce food."

There were few major departures from the conventional lines of thought: As some ancient hunters and gatherers became more specialized and turned to vegetable sources of food as their main diet, agriculture developed and plants became domesticated. The life of these early farmers became sedentary and villages formed. The ability to produce food gave such villagers more control over nature than their forebears had. And with more reliable food sources, population rapidly grew—along with technological innovation and changes in social organization. The ultimate result of all these developments was urban civilization.

The questions of when and where agriculture had its beginnings are still debated among scholars. But the weight of evidence a decade ago seemed to point to the Levant—the Near East—as the cradle of modern civilization. The earliest sedentary villages discovered by archeologists were found there, villages dating back 9,000 to 11,000 years. The excavation of domesticated grain seeds, crude farming instruments, and food-storage facilities indicated that these communities were organized around agriculture. And since the wild relatives of the grains grow in the Levant today, it was logical to assume that this was a natural area for the first food production to take place.

Archeologists generally concluded that agriculture was not an invention as such. Rather, it developed over a period of time as a dominant way of life. Why it evolved, however, remained (and still remains) an open question. This is especially so since research during the 1960s seemed to show that primitive hunting and gathering was, in many cases, a far less demanding and time-consuming existence than farming.

The explanations

British archeologist V. Gordon Childe framed the debate in the 1930s with what is called his oasis theory. Childe believed that the ending of the Ice Age in Europe, some 12,000 years ago, made the climate of the Near East much hotter and drier. As a result, both nomadic peoples and the animals of the region gravitated to the remaining sources of water—rivers and oases. Here too were the surviving stands of wild vegetation. Given the close proximity, people observed the natural cycle of the plants and began to utilize the dense stands of wild grain that grew in the annually watered flood plains. Eventually they started to manipulate these grains, planting seeds as well as harvesting and processing. So, reasoned Childe, began domesticated agriculture—the start of what he called the Neolithic revolution.

Other researchers set out along similar roads. During the 1950s, University of Chicago archeologist Robert Braidwood developed another view: the nuclear zone theory. Braidwood, contrary to Childe, thought that the climate of the Near East had remained virtually unchanged over the past 12,000 years. Thus the beginnings of agriculture would have taken place in a region where wild barley and wheat are still found today—an area that is not too hot or too cold, too wet or too dry. The area on which he focused covers the foothills of the Zagros Mountains in northern Iraq, the so-called hilly flanks of the Fertile Crescent.

Braidwood reasoned that this natural Near Eastern habitat for wild grains and animals attracted nomadic peoples who developed steadily improving ways to use the available food sources. Over time, other settled communities grew up, and the inhabitants came to rely on certain species of wild plants and animals. Later, there would have been experiments with domestication—perhaps the sowing and cultivating of cereals and the husbandry of animals.

In some cases, this experimentation in what Braidwood termed incipient agriculture led to established farming villages and full-fledged agriculture. Evidence at sites dating back nearly 9,000 years appeared to bolster his theory; further research led others to amend this hypothesis. Some scholars became convinced that it might have been population pressures that moved Stone Age peoples of the Near East to gather in villages and turn to farming.

The growth of populations, according to archeologists such as Philip Smith of the University of Montreal, would have taken place in the favorable surroundings of Braidwood's nuclear zone, as more food was needed to support expanding populations, abundant stands of wild plants would have increased in importance. Thus existence would have become more sedentary, stimulating further population growth and a still greater dependence on nearby food. One way of satisfying this need was to learn to cultivate cereals and thus increase artificially the density and distribution of edible wild plants. This cycle led to farming innovations, more intensive cultivation, and finally to full-scale agriculture.

Then, during the 1960s, other scientists refined the population-growth view. Uni-

Excavation. Members of a multinational archeological team explored a number of sites in the deserts of the upper Nile River Valley where controversial evidence of early agriculture by nomadic tribes was uncovered. Dan Grant

versity of New Mexico archeologist Lewis Binford, for example, developed what is called the marginal-zone hypothesis, which expanded on the work of the population-pressure theorists. Binford reasoned that population pressure in food-rich areas such as Braidwood's hilly flanks resulted in the migration of some people to outlying areas where food was less available. To increase the productivity of these marginal areas, the inhabitants brought in plants and animals from the richer nuclear zone. And to make the crops thrive, it was necessary for the people to control and manipulate wild plants. The ultimate result was agriculture and farming villages in these marginal fringes.

The first pass

Wendorf recalls that back in 1966 he had "begun to feel there was something wrong" with these various theories of the origins of agriculture. Nevertheless, he continued to try to reconcile his discovery of grinding stones and crude scythes in the upper Nile Valley with the prevailing ideas.

It was during the following year, 1967, that Wendorf and his colleagues first saw Wadi Kubbaniya. The ancient dry river bed was only some 30 minutes travel time from the city of Aswan, close to a new paved road, and about three kilometers from the Nile. "It was the first day of our survey when we drove into the wadi," he recalls. "There were numerous sites and lots of grinding stones everywhere. I said to myself, 'I don't want to stop now to excavate these sites. There's no threat to them. When I get old and tired of living in tents, I'll come back and excavate them, and live in a hotel in Aswan.'" He decided to go north to other sites.

Farther down the Nile, Wendorf's group excavated what he calls a substantial series of sites at a place called Isna. Scattered around were many grinding stones. And next to one 12,500-year-old site was what once had been a pond. Pollen recovered from the fossil pond was identified as barley. Wendorf calls this "the first clue to what those grinding stones had been used for." In publishing the findings, he notes, "we said we think it deserves consideration that we're dealing here with food production, as far as barley is concerned."

The speculation was greeted with skepticism. There were suggestions, for instance, that the spores had been incorrectly identified. Wendorf's group had no chance to gather further evidence at the time. The Arab-Israeli war erupted and the upper Nile Valley became a military security area. Twelve years passed before he and his coworkers were able to explore the region further. In the interim, he worked in Egypt's Western Desert.

Seeds of heresy

In 1978, Wendorf and his Polish colleague Romuald Schild finally got permission to take their multinational expedition back to Wadi Kubbaniya. The paved road from Aswan had been virtually destroyed by military vehicles, but the sites themselves were untouched. "Some were partially deflated, destroyed by wind action," recalls Wendorf, "but the ground was almost paved with paleolithic artifacts—fossil bones, stone objects, and burned areas."

By this time, Wendorf and Schild had developed a model for what had gone on at Wadi Kubbaniya. "We'd proposed," he explains, "that manipulation of grasses, possibly barley, had begun in the Nile Valley. And it well could have been domesticated. This would have been thousands of years before any known manipulation went on in the Levant." Moreover, the upper Nile Valley was (and is) one of the most arid regions on earth. Since wild barley could not grow there naturally, Wendorf and Schild were convinced that barley grains present at the sites had to have been introduced.

As the days went on, the group painstakingly excavated and surveyed trenches in different areas of the wadi. They searched in the higher dunes adjacent to the floodplain, on the ridges separating depressions in the embayment, and at the mouth of the old river bed. It was clear that the ancient campsites had been repeatedly used.

The largest site had 22 occupation layers, separated from one another by sand. The archeologists discovered fossil remains of Nile catfish, hartebeest, and wild cattle, as well as migrant waterfowl. In addition, there were stone tools and charcoal from ancient hearths. The inhabitants had been small groups of nomadic hunters and gatherers living on the fringes of the wadi some 16,000 to 18,000 years ago.

In the higher sites, located in a dune area some six meters above the old river bottom, there were the intriguing grinding stones and mortars but no trace of seed. A key problem, explains Wendorf, is that in the open character of such late Paleolithic sites—campgrounds with no permanent structures—"everything is against preservation of seeds." If they're not eaten by rodents, birds, and insects, they're washed away by annual floods or tumbled and polished by silt beyond recognition.

"For a seed to be preserved," Wendorf explains, "it has to have been lightly charred so that it's carbonized, but the structure remains intact. Then it has to have been rapidly buried. Too much heat and seeds will shrivel or burn. Thus survival of any grain requires a lucky accident of preservation.

On the next-to-last day of the dig, no botanical evidence had yet turned up. Wendorf directed one of his graduate students to take two samples from a trench in Site Four, one of the upper dune sites. One sample was for carbon dating, the other for botanist Nabil el Hadidi. An hour later, the assistant returned with two plastic-bagged samples. Both had been extracted from a thin layer of charcoal in the trench sidewall. "We took one sample with us and left the other with Hadidi," says Wendorf. "I didn't believe we'd found any cereals."

But Wendorf was wrong. Hadidi's letter, which reached him several weeks later at Southern Methodist University, relayed the exciting news. Four seeds, which Hadidi identified as barley, had indeed been found in the Site Four sample left with him. When the seeds were analyzed later by Ann Stemler and Richard Falk at the University of California, Davis, using electron microscopy, one turned out to be einkorn wheat, which grows wild today some 1,000 kilometers distant from Wadi Kubbaniya.

Lifestyle

On the basis of the seeds, artifacts, and animal remains, Wendorf and Schild tried to envision the lifestyle of the Wadi Kubbaniya dwellers. As the archeologists reconstructed the 18,000-year-old annual cycle, the nomadic bands moved from area to area with the changing of the seasons and shifting sources of food.

The upper dune sites were apparently used by the inhabitants during or just after the annual Nile summer floods, in late August and September. There they captured fish that had been trapped in pondlike depressions among the dunes after the river level dropped. Once the ponds dried, the bottoms of the depressions were left covered by a layer of rich, damp silt—an ideal place to plant legumes and cereals such as barley and wheat.

When the Nile receded further, the bottom of the wadi also began to dry out, and bands of people moved down from the dunes to ridges on the floodplain. There they hunted game and trapped fish from water-filled depressions on the wadi floor. The archeologists believe no planting was done on the wadi bottom, since they found no grinding stones at these sites. Moreover, they reasoned, while the dwellers could use planting sticks in the soft, silty surface of the dunes, their crude farming technology could not cope with the heavy clays of the wadi floor.

Wendorf believes that after the fish supply in the floodplain was exhausted, the inhabitants left for other areas along the Nile, then came back to harvest and process the cereals in the high dune sites during December or January. Remains of ducks and geese, birds that winter in southern Egypt, provided further evidence that the groups occupied the dunes during the winter harvest months.

Wendorf's and Schild's interpretation of what they and their coworkers did not find is as provocative as their explanation of what they did uncover: Nothing was discovered to indicate that wild cereals grew naturally in the area, nothing to show any environmental stress, nothing to point to population pressure. In other words, none of the theories which sought to explain people's Stone-Age motivation to farm seemed to apply to the people of Wadi Kubbaniya. Moreover, there was none of the subsequent population growth or social and economic change thought to flow inevitably and swiftly from the development of food production.

In short, Wendorf and Schild are convinced that the rudimentary agriculture that appears to have been practiced at Wadi Kubbaniya and elsewhere in the Nile Valley some 12,000 to 18,000 years ago simply provided another food source in what they term a "broadly based hunting, fishing, and gathering economy."

These new findings and interpretations met with a skeptical reaction. The discovery of cereal seeds so far from their known natural ranges was indeed seen as significant. But some scientists wondered whether the site had been somehow contaminated, perhaps penetrated by material washed in from a stratum laid down during a considerably later period. A few even suggested that a graduate student might have planted the seeds in the sample as a prank. Other archeologists questioned whether four cereal grains constituted a sufficient sample.

Domestic or wild

The key point raised, however, involved the seeds themselves. Although one seed and its stalk attachment indicated domestication, the others lacked such attachments. And it could not be determined whether they were from domesticated or wild plants. The distinction is important: When the seeds of wild grains are ripe, their attachments to the stalk decay. The grains then fall off cleanly and easily, with any movement, and are widely scattered. This is the mechanism by which wild grains propagate.

As mutant plants arise, however, some emerge with seeds that are more firmly attached to the stalk. When ancient gatherers collected grain, these would be the plants that remained to be harvested. And when the earliest farmers began to practice agri-

---

lker is a former Time *correspondent who writes frequently about environmental and ecological subjects.*

culture, the grains stored for the next year's sowing tended to be from such mutant plants.

Over time, the selection process built on this genetic difference; more of the cereals planted and harvested were those with firm attachments. When the plants became truly domesticated, none of the ripened seeds fell off naturally. They had to be forcibly torn away from the stalks; threshing was required. The difference between seeds of wild and domesticated plants can be seen by electron microscopy: If the seed is wild, the attachment to the stalk will be smooth; if the seed is domesticated, the attachment portion will be torn and jagged.

Since only one of the four cereal grains recovered by the expedition had the jagged stalk attachment that indicated domestication, the evidence was inconclusive. And skeptics claimed that the dwellers of the wadi were simply gathering wild grain rather than practicing agriculture.

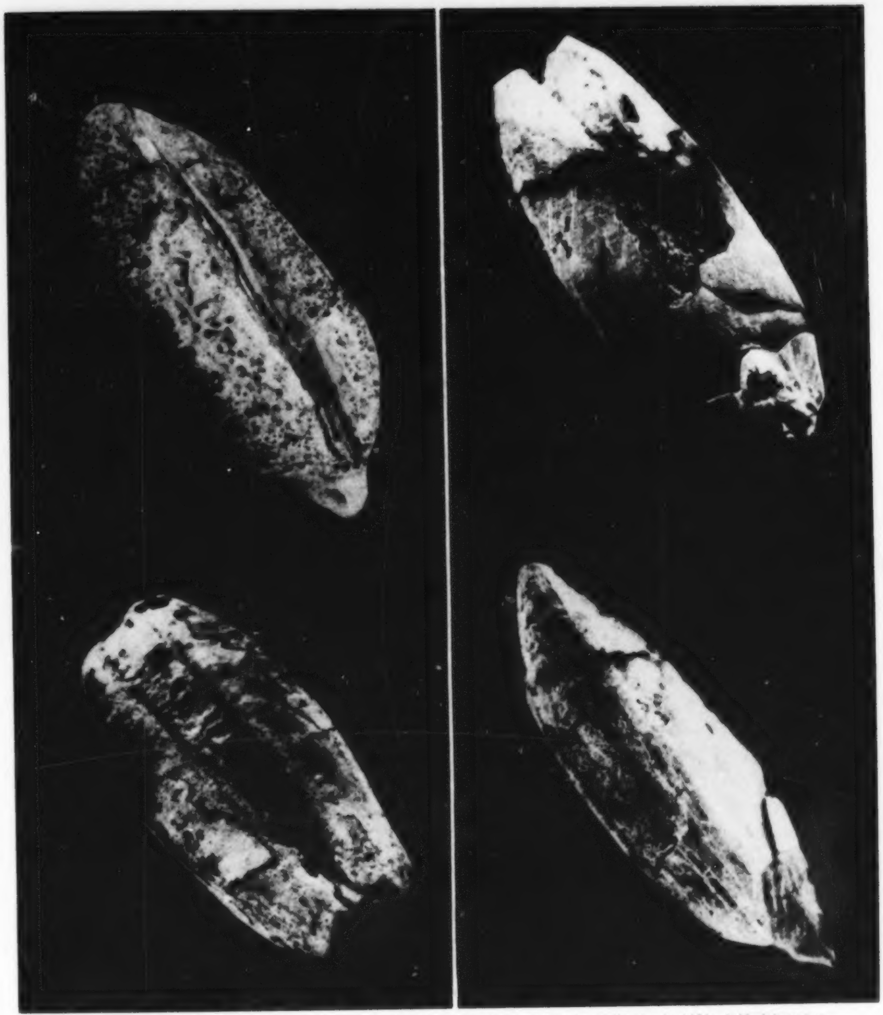
In 1980, the joint United States-Egyptian-Polish expedition returned to Wadi Kubbaniya to search for more conclusive evidence. "This time we had three botanists with us," Wendorf says. "We searched the site where we'd found the seeds for days without finding an additional grain of cereal. At that stage we were asking ourselves whether we were wrong. Talk about depression..."

By the beginning of 1981, declares Wendorf, "we devised a new screening technique which involved putting every bit of earth we excavated through not just one screen but through a series of graded screens." The elaborate sifting process was to make certain the scientists wouldn't overlook even the smallest bit of botanical evidence.

It was two weeks before the search paid off. Southern Methodist archeologist Angela Close, who is Wendorf's wife, was working in Site Three, another high-dunes campground. The excavated trench showed 22 successive cultural layers. Material she had removed from the twenty-first layer was placed on the top screen and, she recalls, "suddenly we saw cereal grains." By the time she was finished, Close had found over 200 grains of barley. And from elsewhere in the dig, lentils, chickpeas, and date seeds were recovered as well.

Most of the cereal grains were concentrated in two small clusters. "Apparently there was a little ripple in the dune," explains Wendorf, "and the wind carried some seeds from roasting or grinding sites nearby. Then the ripple was covered over by sand. It was one of those accidents."

The strata from which the rich finds were excavated were later dated as ranging from 17,800 to 18,450 years old. The discovery proved that the original find was no fluke. The sizable cache of grain seemed to bolster Wendorf's hypothesis, but not nail it shut. None of the seeds had the stalk attachments necessary to determine whether they came from wild or domesticated plants. This key question remained unresolved.



Specimens. Grains of barley (left pair) and wheat, part of the sample collected at Wadi Kubbaniya, provided the first tangible support for the contention that primitive agriculture began there 18,000 years ago.

Ann Stemler, Richard Falk/University of California, Davis

Colleagues contend

Archeologists who have long studied the origins of agriculture in the Mideast agree that Wendorf's latest evidence from Wadi Kubbaniya is important. But many feel the significance lies in what they see as new knowledge about the uncertain geographic distribution of wild cereals during the late Pleistocene period. At the same time, however, many still question the Wendorf-Schild contention that primitive agriculture was being practiced in the upper Nile Valley some 18,000 years ago.

These scientists believe that the grain harvested and processed at Wadi Kubbaniya came from wild stands. And they echo University of Montreal archeologist Philip Smith, who agrees that "people there were extensively using wild plants. But this is what we expected all along among primitive people fifteen to twenty thousand years ago. There's nothing very surprising about it."

"People were doing the same thing in Southeast Asia at the same time and in other parts of North Africa and very likely in India," says Smith, one of those who developed the population-growth theory of the origins of agriculture. "So why not in Egypt? People in many parts of the world during that period were getting quite clever, quite inventive. And possibly in one or two places in the world [that development] eventually did lead up to full-scale plant domestication and animal husbandry.

"But it didn't have to. And in most places, of course, it didn't," Smith adds. What was going on at Wadi Kubbaniya, he concludes, "was simply plant exploitation, fairly intensive plant exploitation, but not domestication or even cultivation at this time."

Wendorf maintains that the extensive wild stands of grain needed for such intensive exploitation could not have grown in the upper Nile region at the time because of the extreme aridity and heat. Other scientists, however, question whether that's necessarily the case. Geologist Herbert Wright of the University of Minnesota, for example, theorizes on the basis of pollen research that the Maghreb region to the north was cooler and wetter during the late Pleistocene than it is today. And he speculates that North Africa might have been a refuge for such cereals as wild barley and wheat.

University of Illinois plant geneticist Jack Harlan adds that the presence of barley in the upper Nile "certainly implies a more radical change in climate than we'd previously supposed, a more radical shift in the distribution of these wheats and barleys. But it's not really unexpected. We knew there was a change. After all, most of Europe was still under ice during this period."

Harlan and others don't dispute Wendorf's claim that the upper Nile Valley was extremely dry some 18,000 years ago. But they feel that the annual Nile flooding would have provided sufficient moisture to support large stands of wild cereals if the climate then were indeed cooler. "We're dealing with an area that's part of the Nile floodplain or would be affected by seepage of water from the Nile on a seasonal basis," geoarcheologist Karl Butzer of the University of Chicago maintains. "In either case, you'd have seasonal flooding which would allow wild plants of that general nature to grow without any trouble."

Why don't the cereals grow wild in the same area today? Butzer notes evidence of a worldwide warming trend that occurred 8,000 to 12,000 years ago. (See "Climate: How Large an Unknown in the Food Equation?" *Mosaic*, Volume 6, Number 3, and "Climate, Weather, Aridity," *Mosaic*, Volume 8, Number 1.) His point: Temperatures in the region over the past 10,000 years or so may be too high to allow wild grasses like barley and wheat to compete successfully with other plants.

Acceptance's spectrum

University of New Mexico archeologist Lewis Binford, who developed the marginal-zone demographic theory of agriculture's origins, makes some concessions to Wendorf, but he questions the breadth of Wendorf's interpretations: They would not, he says, nullify existing theories of the origins of agriculture. "What [Wendorf] has is another case," declares Binford, "presumably with some variability. So it has to be dealt with as a case, not as a testing of hypotheses, because the patterns of variability are so complex."

Another population-pressure theorist, Mark Cohen of the State University of New York at Plattsburgh, feels that what Wendorf's findings demolish, if anything, "is sort of an old-fashioned notion that somebody discovered how to plant seeds, and from that you must get a Neolithic revolution, including centralization, growing populations, and so on. Clearly his work demolishes that. It divorces the event—some people trying this as an experiment—from the later commitment."

Indeed, says Cohen, "what I think Wendorf is saying, and I agree with him, is that there's a great difference between people practicing and perfecting their technique on something that doesn't become an economic staple and their making a commitment to sedentariness and a major dietary system. I'm primarily concerned with this much more widespread and irreversible trend that came later. I'm less concerned with determining when or where any species was first domesticated."

Charles Redman, an archeologist at the State University of New York at Binghamton, sees greater significance in Wendorf's findings. If the peoples of Wadi Kubbaniya were manipulating plants, he argues, "what it does is change the context for me in thinking about the problem. I consider that a basic thing for us—to rethink whether it's agriculture itself that had such a tremendous impact on society, or whether agriculture had great impact only in certain cultural/environmental contexts and was a relatively unimportant innovation in others." Perhaps, Redman continues, "we have to reevaluate some of these prime movers we've been looking at. I think what we're finding—and Fred's evidence may push it further—is that we're looking at much more of a mosaic in the Mideast than we thought, and there could be much more variation than we believed."

Even if there are cases in which early agriculture proved to be what Wendorf termed a nonevent, Redman says, "the same event, or whatever you want to call the origins of agriculture, did have a profound impact in many other places. So the point is that the invention, or the use of plants, isn't in itself what transforms society. But in certain circumstances, it was part of the process that did."

Botanical continuity

Although the reaction to Wendorf's findings and interpretations is mixed, he remains unfazed by the skepticism. For example, a critical point of debate is whether the peoples of Wadi Kubbaniya were practicing crude agriculture or merely exploiting wild plants. Wendorf continues to maintain that the necessary abundant wild stands of cereals didn't exist in the region. As proof, he points to the widespread fossil evidence found at the sites that reveals the native wild flora present there 18,000 years ago. "It's the same vegetation," he says, "that survives today on islands in the Nile. And you don't find even feral stands of barley and wheat trying to compete with these native plants, bushes, and trees." In other words, the wild plants of the area 18,000 years ago are the same as those surviving now—and there's no trace of wheat or barley growing wild during either period.

Moreover, he notes, the cereals and legumes must have been growing in lower areas among the high dunes sites some five to six meters above the floor of the wadi. The annual Nile floods vary enormously in intensity, he explains; thus the sites in the dunes could remain high and dry for several years on end. And, Wendorf adds, "cereals won't hibernate over such long periods without water. They won't germinate two or three years later.... We can't absolutely prove that wild cereals didn't grow in the Nile Valley somewhere. But we can document that it's highly unlikely at Wadi Kubbaniya."

Both Wendorf and his critics agree that the indisputable proof of his theory would be the discovery of a large number of seeds bearing ragged attachments—grain demonstrating that the genetic change accompanying domestication had taken place. Wendorf believes that change did occur. "But we have to find a large sample," he says. "Maybe we'll get it, maybe we won't. If we do, we can resolve the question."

However, he insists, whether the seeds unearthed at Wadi Kubbaniya prove to be wild or domesticated is not critical. He firmly believes that the plants would not have existed at the sites unless they were introduced and manipulated by humans. Whether the seeds are wild or domestic, he argues, the people there were dealing with them as though they were domestic. Primitive agriculture was being practiced at Wadi Kubbaniya, and to

Sifting sand. Every bit of earth excavated at Wadi Kubbaniya was sifted through a series of screens to make certain that even the smallest bit of evidence would not be overlooked.

Clint Grant

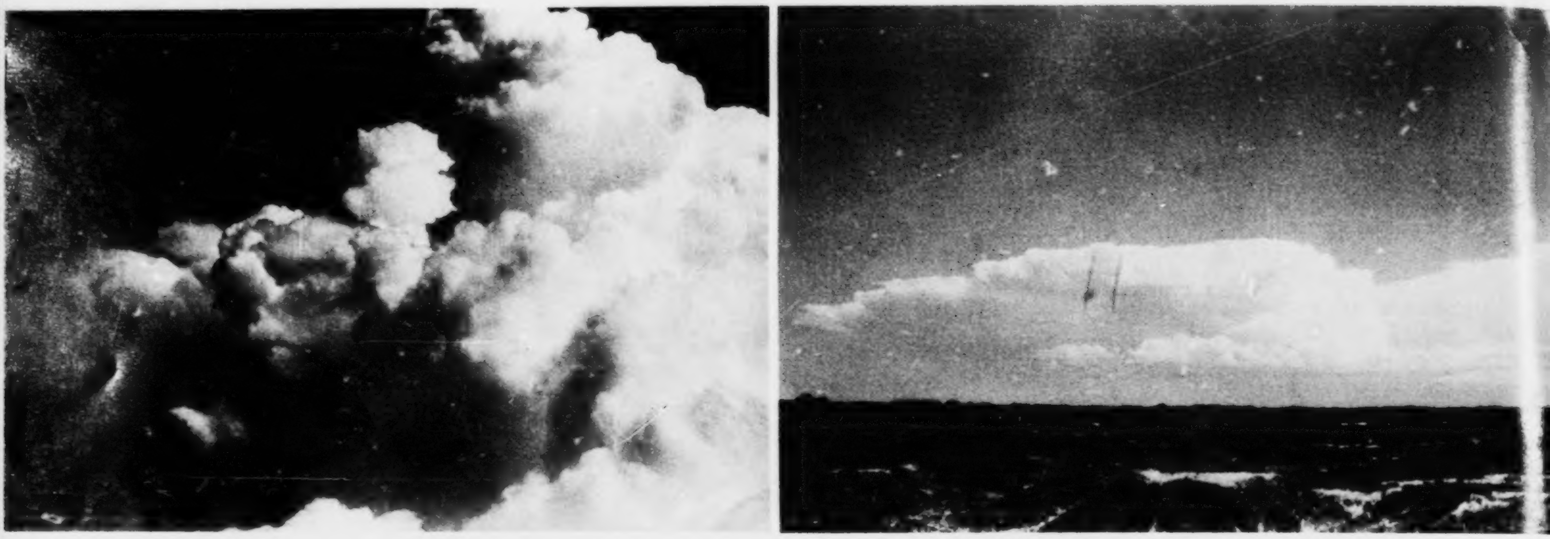
Wendorf this is more important than genetic changes in the plants themselves.

Moreover, given the already developed grain processing and harvesting tools of the time, Wendorf believes this sort of rudimentary farming probably had begun far earlier in the Paleolithic period than 18,000 years ago.

"We've been looking at the wrong thing," he says. "It would seem that what needs to be explained is not the beginnings of food production; I think that was a normal and natural response to new food resources. What we need to explain is the shift in emphasis to dependence on cereals—the increased emphasis on cereals as a source of food which ultimately led to the Neolithic revolution."

Meanwhile, those who remain skeptical about Wendorf's theories continue to feel that, as Chicago's Robert Braidwood puts it, "the jury is still out." Adds Braidwood, who developed that influential nuclear zone theory of agriculture's origins in the 1950s, "If Fred gets more evidence, and there's general agreement that there was indeed positive manipulation there, then bless his heart." •

The National Science Foundation contributes to the support of research discussed in this article through its Anthropology Program.



Evolution. Growing cumulus congestus clouds (left) develop into cumulo nimbus anvils as they reach the tropopause. Snow crystals collected in a cumulus congestus appear like the replica (right).

Ice in the

by Henry Lansford

In a massive, coordinated effort, scientists are exploring the dynamics of summer storms. How ice forms in storm clouds and in turn influences cloud formation is high on the list of questions.

Ice crystals form and multiply in the frigid upper levels of summer thunderstorms. Many things help determine how much and in what form precipitation will fall from the storms, a familiar sight over the earth's tropical and temperate regions. They occur on an average of 100 days a year over South Florida and 70 days a year over the Colorado Rockies. Some atmospheric scientists estimate that three or four thousand thunderstorms are in progress around the world at any given moment.

Many thunderstorms are potential killers, spawning lightning, hail, flash floods, and tornadoes. But they also bring rain to areas such as the North American breadbasket, the great food-producing agricultural region of the central United States and Canada. If spring and summer thunderstorms fail to bring enough growing-season moisture to such regions, drought can wipe out millions of acres of vital food crops.

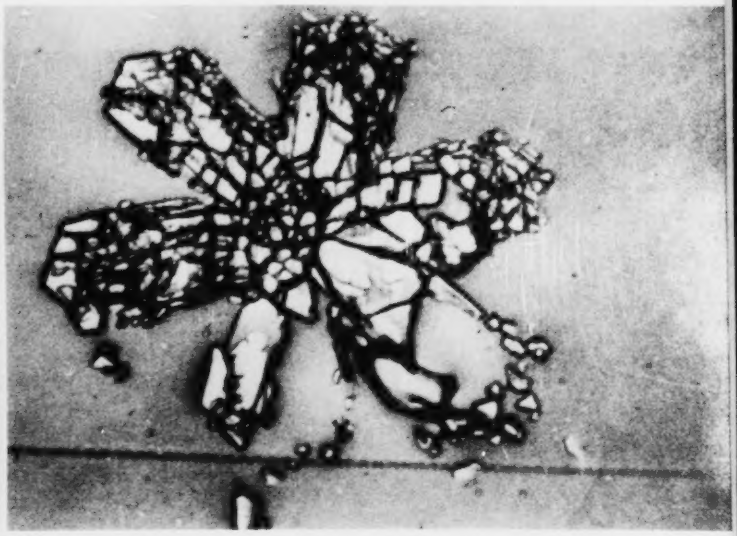
Many seeding attempts have been made to modify thunderstorms, to try to increase rainfall or decrease crop destruction by hail from these storms. Some farmers and commercial cloud seeders are convinced that their weather-modification projects have been effective. But most atmospheric scientists believe that a much better understanding of the physics and dynamics of thunderstorms is needed to assess realistically the effectiveness of such efforts.

The process that often starts thunderstorm machinery is convection—the upward movement of parcels of warm, moist air from near the ground. This movement can set off a virtual explosion of energy as latent heat is released when water vapor in the cooling, expanding air changes state, forming liquid droplets and ice crystals. This heat energy accelerates the upward motion of the air, and the storm grows deeper. If the atmosphere is sufficiently moist and unstable, a small cumulus cloud can grow into a thunderhead ten kilometers high and ten across in a matter of minutes.

Alone or together

Some thunderstorms occur in solitary splendor, mushrooming up against a brilliant summer sky. Known to meteorologists as an isolated, or air-mass, thunderstorm, this kind can unleash as much energy during its brief life as is released by an atomic bomb of moderate size.

Isolated thunderstorms are not rare, says research meteorologist Douglas Lilly of the National Center for Atmospheric Research in Boulder, Colorado. But "more often they occur in crowds." Ranks of thunderstorms



Summer Sky

often form along an advancing cold front, triggered by a wedge of cold air that forces warm, moist air upward along the front, starting the convective process.

Whether the initial upward motion of air is triggered by a cold front or local convection over a plowed field, many of the tiny droplets and crystals that form in a convective cloud can grow into larger particles—raindrops, snowflakes, hailstones, or snow pellets called graupel. If these particles grow large enough, their fall velocity will exceed the speed of the storm's updraft. It will, as they say, precipitate. On hot summer days, the snowflakes, graupel particles, and small hailstones will melt on the way down and will reach the ground as raindrops. Large hailstones normally are the only solid precipitation particles that fall to the earth from summer thunderstorms.

Although atmospheric scientists agree on some broad concepts that explain the formation of precipitation in convective storms, there is uncertainty and some controversy. Attempts to predict the amount, kind, and rate of precipitation that will result from a given set of initial conditions have not been very successful. The precipitation processes that occur in warm clouds, where no ice is present, are better understood than the chain that produces precipitation in cold clouds, says William Cooper, a cloud physicist with the University of Wyoming. And it is the cold-cloud processes—with some emphasis on ice formation—that are attracting the most research attention.

Ice formation

"There have been efforts over more than 20 years," Cooper says, "to measure characteristics of the air that will predict how much ice is going to be present. Those predictions are still quite unsuccessful. We try to measure concentrations of ice nuclei in the air and then relate these to the amount of ice that we actually find in the clouds. We usually predict far less ice than we actually find."

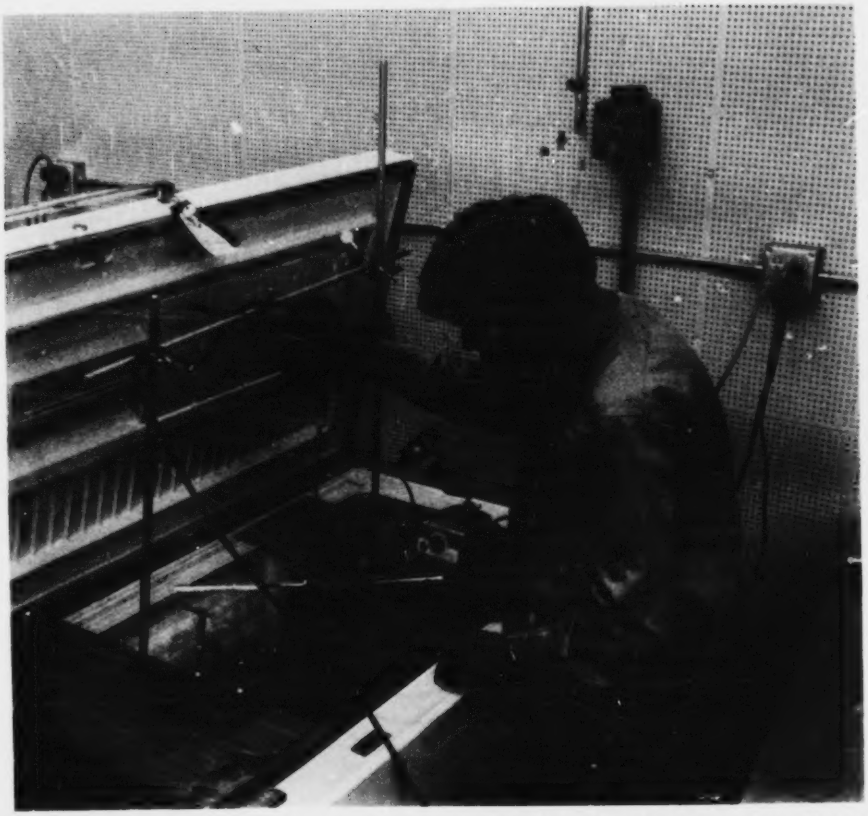
Cooper says one complicating factor is that ice-nucleus measurements do not seem to be representative of even the initial ice formation; a second is that ice-multiplication processes produce more particles than can be accounted for by the original ice nuclei. One such mechanism is known as the Hallett-Mossop secondary ice-crystal production process. It proposed that additional ice particles are produced by graupel—the snow pellets that form by accretion as ice crystals collide with supercooled cloud droplets—and then fragment.

"It turned out not to be quite that simple," says one of the originators of this hypothesis, John Hallett of the University of Nevada's Desert Research Institute. "If indeed this process was taking place, it had to be under rather special conditions." Hallett and his colleague Stanley Mossop, of the Commonwealth Scientific and Industrial Research Organization in Sydney, Australia, decided that such conditions would most probably occur when the original ice crystals grew from the graupel in a fragile configuration that would be most subject to shattering on impact with a droplet.

"Snow crystals come in a variety of shapes," Hallett points out. "You have long, thin things that look like pencils, you have six-pointed, Christmas-card dendrites, you have flat plates, and you have all sorts of combinations."

Some of these configurations are more fragile than others; moreover, the shapes offer clues to the conditions under which the snow crystal grew. Thus the conditions appropriate to secondary crystal development can be inferred.

Hallett argues that if the breakup of vapor-grown crystals were providing secondary particles, it was probably occurring at temperatures near either four degrees or fifteen



Growing crystals. John Hallet adjusts the rotating arm of a device used to grow ice crystals as part of the research on secondary ice-crystal production in convective storm clouds.

Henry Lansford

degrees below zero centigrade. At minus four, crystals grow as long, thin needles; at minus fifteen they form delicate dendrites. Both are fragile. To test this hypothesis, Hallett and Mossop set up a laboratory experiment, growing graupel on a rotating arm in a cold box filled with cloudy air.

"We tried the lower temperature first, and we couldn't see anything. Then we tried it at minus four degrees centigrade, and there were crystals, just like that. We tried it at various temperatures and showed that this particular multiplication process takes place only between about minus three and minus eight degrees centigrade."

It has since turned out that the original reason the scientists thought the crystals formed "probably isn't right," Hallett says. "It seems to have something to do with the freezing of the drops. But nevertheless it... established that the process occurs only over a particular temperature range." The laboratory experiments also established another factor that was necessary for the process to operate—a particular cloud-drop spectrum: "...a fairly large number of very small cloud drops and a moderate number of rather large cloud drops," Hallett says.

Hallett then set out to confirm that the process would actually produce secondary ice crystals when the necessary conditions occurred in the real atmosphere. Using a NOAA research aircraft, he penetrated convective clouds and found high ice-crystal concentrations accompanied by the temperatures and cloud-drop spectrum that had been present in the laboratory experiments. "To cut a long story short," Hallet says, "the field work showed that the concentrations of ice were consistent with the secondary ice-crystal production hypothesis that we had obtained from the laboratory work in Sydney some four or five years earlier."

A simple model

Dennis Lamb, another colleague of Hallett's who is also at the Desert Research Institute, has developed a numerical model designed to quantify the ice-production mechanism. "Our modeling," says Lamb, "is at a very fundamental level, one that some modelers would consider simplistic. We are trying to get at the minimal set of mechanisms that can explain the data, even though there clearly are a lot of other things going on."

Lamb's model involves large and small cloud particles either as liquid water or ice. The model starts out with no ice at all, he says, just a given distribution of cloud droplets and raindrops and a certain concentration of nuclei—airborne particles that give condensation a place to start. (See "A known mechanism," accompanying this article.)

The initial ice formation is either by contact nucleation, in which the nucleus touches a supercooled cloud droplet or raindrop, or by deposition nucleation, in which water vapor is deposited on the nucleus as ice crystals. In either case, there is a one-to-one relationship between the nuclei and the concentration of these primary ice particles in the cloud. And that relationship holds, despite the fact that ice particles—secondaries—continue to form.

What seems to happen, Lamb says, is that at some point tiny ice splinters are thrown off. They become instrumental in freezing and accreting large, supercooled drops in the cloud, and secondary ice production begins.

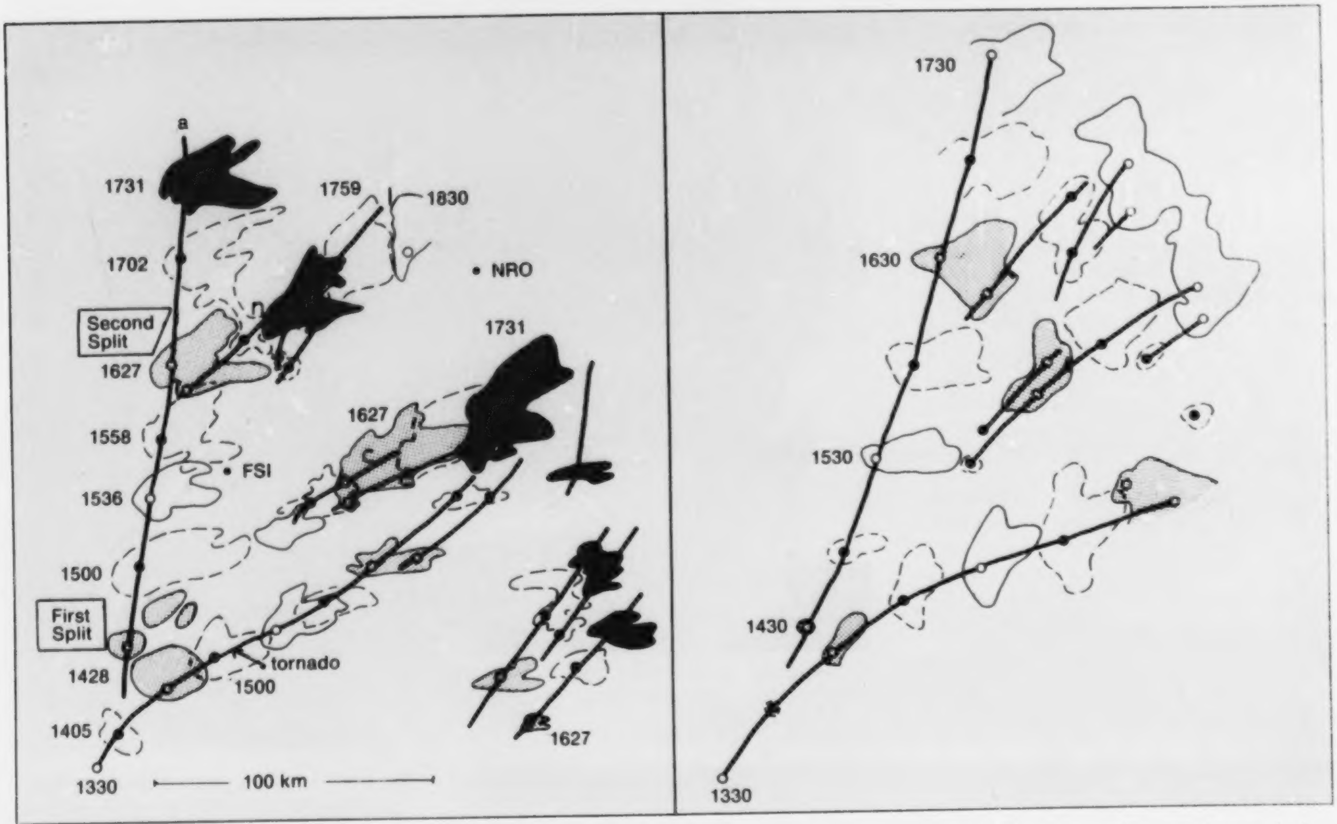
With the model, Lamb points out, he can do something that is impossible in the laboratory or real atmosphere: compare the process with and without the secondary mechanism. If secondary production is eliminated, the ratio of nuclei to ice particles remains one to one. If it is included, says Lamb, "We have a nice exponential increase in the ice-particle population until all the raindrops are exhausted. This would account for the high ratio of ice to ice nuclei that is found in many clouds."

Other scientists have criticized this approach as oversimplified, a criticism Lamb accepts. "One thing that is distinctly lacking in this model," he says, "is that it is strictly microphysical, without any dynamical features. We know that the particles are going up and down with the updrafts in the clouds and that we are not taking into account a lot of things that are going on in reality. But this model does allow us to explore the interactions of certain types of ice-producing mechanisms in some detail."

Dynamic models

Other modelers have taken a different approach, concentrating on dynamics—the air motions in thunderstorm clouds—rather than the microphysical interactions of tiny water droplets and ice particles.

Lansford is former public affairs officer of the National Center for Atmospheric Research and coauthor of "The Climate Mandate."



Storm simulation. A computer-generated model simulating the process by which a severe thunderstorm splits into two separate storms (left) is compared with observations of the actual event. Such simulations are still more a research than a predictive tool.

Robert Wilhelmson and Joseph Klemp

"We began our study by assuming that it is not necessary to represent the detailed microphysical interactions in a numerical cloud model in order to describe the airflow within many storms," says Robert Wilhelmson of the University of Illinois. Wilhelmson has collaborated with Joseph Klemp of the National Center for Atmospheric Research to demonstrate the validity of this assumption. They have done several studies that used a numerical model to describe the airflow in all three spatial dimensions and in time, producing over the past five years what Wilhelmson describes as important advances in the understanding of storm behavior.

In one study, Wilhelmson and Klemp investigated the dynamics involved in the splitting of a single thunderstorm into two separate storms and eventually into a line of storms. Their modeled pattern of storm development compared favorably with the pattern obtained from radar observations of such an event that occurred in Oklahoma.

In another study, done in cooperation with Peter Ray of the National Severe Storms Laboratory in Norman, Oklahoma, Wilhelmson investigated the airflow structure of a tornadic storm by comparing the flow fields in an actual storm observed with Doppler radar to modeled flow fields. The model storm was generated by embedding a small cloud in a simulated environment representative of the conditions under which the observed storm formed. This simulated cloud and its associated updraft grew in height and intensity. Precipitation then formed along with a downdraft. The main updraft and downdraft in such storms complement each other, with precipitation being displaced out of the rotating updraft into the downdraft, and such a storm flow structure can remain relatively constant for some time.

The model flow fields were quite similar to those observed just before a tornado formed in the observed storm. Because of this, Wilhelmson and his colleagues are confident that the model can be used to fill gaps in observation of wind fields and to study the interaction between various dynamical processes that cannot be evaluated with the wind fields alone.

Wilhelmson emphasizes that such dynamical studies have been done only by making certain simplifying assumptions concerning detailed microphysical and turbulent processes that occur on scales smaller than the model's resolving power—about half a cubic kilometer. Sufficient computational power just doesn't exist to model all the processes that occur within a storm on scales ranging from microns to kilometers, he observes, yet "we can gain insight and understanding of storm development and structure by making certain approximations."

Numbers to crunch

One limiting factor on improvement of both microphysical and dynamic models is a shortage of data. The most pressing need in thunderstorm research has always been for accurate, comprehensive measurements inside and around the cloud throughout all the stages of its growth and decay. But taking measurements inside a thunderstorm is not an easy task.

Scientists have been penetrating thunderstorm clouds with instrumented aircraft since 1947, when Horace Byers of the University of Chicago did it with a fleet of ten World War II P-61 Black Widows. This was the first major scientific field program designed to probe and analyze convective storms. Recent advances in instrumentation technology have enabled aircraft to make accurate *in situ* measurements of important aspects of the storms. But these measurements are still far from comprehensive.

As Patrick Squires of the National Center for Atmospheric Research puts it, "Aircraft provide reasonably clear-cut measurements of temperature, liquid water content, and updraft. But they are at best line measurements along the flight path through the cloud. It isn't good enough just to fly through the cloud and say that's the way it is, because the situation will be quite different three minutes later or half a mile to one side or the other."

One way to improve this view of a storm is to penetrate the cloud simultaneously with aircraft at different levels. But if the cloud is ten kilometers high and ten kilometers across, even a dozen sets of simultaneous aircraft measurements will still give a badly fragmented picture of the storm.

Coordinating. Brant Foote, of the National Center for Atmospheric Research, was among the coordinators of the land-based and airborne research effort at the CCOPE operations center.

Henry Lansford

Another observing tool, radar, can provide an overall view of a thunderstorm that complements the direct aircraft measurements. And Doppler radar can provide valuable measurements of velocities of particles in the cloud. "This can be translated into the velocity of the air itself, but it's a rather rough-and-ready procedure," says Squires.

A known mechanism

Precipitation is known to form in clouds by two principal mechanisms. In the collision-coalescence process, droplets falling at different velocities collide and merge to form larger droplets. In some warm convective clouds over tropical oceans, rain appears to be produced entirely by collision and coalescence.

But much of the earth's precipitation, particularly over the continental areas where most of our food is produced, forms by an ice-crystal process. Whether it reaches the ground as raindrops, snowflakes, or hailstones, this precipitation starts out as particles of ice in the cloud. The process by which they grow is usually called the Bergeron process, after the Scandinavian meteorologist Tor Bergeron, who performed the definitive studies of it around 1930.

The Bergeron process depends on two fundamental properties of water. First, droplets of pure water suspended in air will not freeze at zero degrees centigrade; they must be chilled to nearly minus forty degrees centigrade before they will freeze by themselves. Water droplets that are still in the liquid state at temperatures below zero degrees centigrade are generally referred to as supercooled. Supercooled droplets will freeze on contact with ice crystals or with particles that have a crystal structure resembling that of ice. Such particles are called freezing nuclei or ice nuclei.

The second property of water that is involved in the Bergeron process concerns its vapor pressure. Air that is saturated with water vapor with respect to liquid water is supersaturated with respect to ice. This means that when ice crystals and water droplets are both suspended in very moist air, water vapor will form additional ice crystals on the particles that already exist and the droplets will evaporate to replace the lost water vapor. Thus the ice crystals will grow rapidly at the expense of the water droplets. As an ice particle grows large enough to precipitate through the cloud, it will continue to grow by colliding with other supercooled droplets that freeze on contact, building the particle by accretion. •

Instrumented aircraft and radar were the two principal observing tools used in the summer of 1981 in the Cooperative Convective Precipitation Experiment, the largest and most ambitious thunderstorm field research program ever mounted. The project, called CCOPE, was planned and run by NCAR and the United States Department of the Interior's Bureau of Reclamation, which has been interested for many years in the potential of weather modification in the West.

The experiment was conducted over an area of eastern Montana, about 300 kilometers across, centered on Miles City. It was a large, coordinated field program that focused a variety of measuring techniques on individual clouds. The aim was to provide data sets that would give an accurate and comprehensive description of convective systems. It would do in one place, in a coordinated way, what individual research groups have been doing alone for a long time.

Patrick Squires, who was codirector of CCOPE along with Bernard Silverman of the Bureau of Reclamation, considers ice forma-

Graupel. Ice particles are responsible for the production of secondary crystals and separation of electrical charge as collision and bounce-off occur within the clouds.

John Hallett

tion, particle movement, and air mixing within clouds among the principal problems addressed. They are all related. The origin of ice in a cloud—at the heart of questions of weather modification by cloud seeding—also concerns the evolution of precipitation particles in ice clouds, he says. Further, to understand what kind of precipitation will fall from a given cloud, "you have to understand the movement of particles between different environments in various parts of the cloud," he declares. And this ties in turn to the third primary objective of the experiment: the effort to understand the process by which dry air from aloft mixes with moist air that has been lifted by convection from layers of the atmosphere near the ground.

"We're trying to understand the cloud as a whole—how ice particles form in various parts of the cloud, how they are transported to other parts, and how those parts get to be the way they are when the particles arrive there," Squires explains. "To do this, we need a four-dimensional view of the cloud, as it were."

Besides that from aircraft and radar, most of the information on the atmospheric environment came from rawinsondes—balloons equipped with sensors to measure temperature and humidity, and a radio transmitter to send the data back. The movement of the balloons was tracked to provide data on winds aloft. Unlike much of the experiment's instrumentation, the rawinsondes were conventional equipment that the National Weather Service uses to make routine observations.

A network of ground-based instruments was also used to measure what was happening beneath the storm. These networks provided measurements as the events occurred, but they were limited by being in fixed locations. As Squires puts it, "They are very important when the cloud you're observing happens to be over the network."

Nerve center

The component of the cooperative experiment that brought all of these observing techniques together, focusing them on the same place at the same time, was the project's operations center. This enabled one scientist—the operations director for the day—to orchestrate the whole ensemble of aircraft, radar, and other observations. As Squires describes it, "That control center was a masterpiece. It gave one capable scientist the ability to be aware of all the information that was coming in and to plan the next 20 minutes or so of operations. Everybody in the field had someone in the center to talk to. If you were operating a Doppler radar, there was a controller in the center to give you your azimuth and elevation and scanning procedures so you were illuminating the same volume of the storm that was being observed by the other radars and penetrated by the aircraft."

The operations center was a noisy and busy place when a research operation was in progress. Controllers seated at consoles might be talking over as many as six radio channels at once—four frequencies were available for communicating with aircraft, one for radar and one for mobile ground units. The heart of the activity was a system known as MAPS, for multiple aircraft positioning system. Three interferometer stations around the network tracked each of up to ten aircraft by triangulation, and an altimeter on each aircraft transmitted its altitude. A color display on the operations director's console showed the position and altitude of each aircraft, along with radar information on the storm itself.

The job of operations director carried a good deal of pressure with it. NCAR's Brant Foote took turns with Charles Knight, also of NCAR, and Arlin Super of the Bureau of Reclamation. Each served for three or four days in a row, then took a break for several days.

Research aircraft. Instruments mounted on a research aircraft are used to measure water and ice particles, air motions, and other aspects of thunderstorm clouds.

Henry Lansford

Foote is a veteran of many hours in the air on research missions, as well as stints as operations director for earlier field work in the National Hail Research Experiment (See "The Clouds With Teeth," *Mosaic*, Volume 3, Number 4). He notes that "I'm not aware of any other experiment that has used this many aircraft operating this close together. On some days, we had three or four different missions going at the same time in different parts of the cloud. We designed some standard missions, but we had to be opportunistic. If nature produced something a little different from what we had written in the book, we had to adapt to it."

As might be expected, the forecasting operation that guided each day's research preparations used the most advanced techniques and data available for predicting convective storms. But the uncertainty of such forecasts was emphasized by the events on one of the project's best research days, July 11, 1981.

Through most of that day, the forecast was not promising for any serious convective action. There were thin bands of clouds over the area, and only one research aircraft was up. But then one band of clouds began to grow more intense, and around 6 p.m. the southern end of one cloud line boiled up into a major thunderstorm. All of the aircraft were called up, and overlapping missions were flown through and around the storm, measuring liquid water content, particle size spectra, temperature, and updraft velocity. "We were ecstatic," Brant Foote says. "The whole thing was really unexpected."

In spite of its chronic unpredictability, the weather generally cooperated with the scientists by providing the kinds of things they wanted to study. The aircraft flew on about six out of every seven days during the three months that the researchers were in the field. As the University of Wyoming's William Cooper describes it, "It would be hard to imagine better cooperation from the weather. We had consistent development of clouds in the area that Dopplers could scan, far more than in any other year. This is the fifth year that I've worked in the area, and I've never seen clouds as suitable for this kind of research as we've had this year."

Preliminary evaluation and analysis of the data began as soon as the field work was over, and a data inventory was available early in 1982. But the results will emerge gradually over a much longer period of time. William Cooper says that the data analysis should be at least as complex and challenging as mounting the field effort. "The proof of how successful we were in designing and carrying out this experiment," he says, "won't come until many years after the field work was finished."

There is no doubt that Cooperative Convective Precipitation Experiment will supply a wealth of data for scientists who work with thunderstorm models. It should also provide weather-modification researchers with valuable fundamental knowledge that may lay the foundation for future cloud-seeding experiments. At some point down the line, severe-storm forecasters should be able to make better predictions of thunderstorm behavior by integrating this new knowledge into their operations.

But valuable though the results may be, there is not a chance that they will suddenly illuminate all of the dark empty spaces of uncertain understanding as to how thunderstorms work. The scientists who study convective storms have their work cut out for them for many years and decades to come. •

The National Science Foundation contributes to the support of the work discussed in this article through its Division of Atmospheric Sciences.

Born to Navigate

by William J. Cromie

Many marine species come armed with senses that enable them to perform feats of navigation for which humans need batteries of sophisticated instruments.

Large bluefin tuna tagged in the Gulf of Mexico have been found off the coast of Norway, more than 8,000 kilometers away.

Pink salmon feed off the coast of Alaska, then travel as much as 7,800 kilometers to spawn near Japan or Korea.

Green turtles regularly migrate from Brazil to remote Ascension Island, a commute of 2,200 kilometers to a 38-square-kilometer spot in the ocean.

Fishermen have known for centuries about such journeys, which involve months of travel across oceans remarkable for their apparent lack of trails and guideposts. Not until the last few years, however, did researchers learn that tuna, sea turtles, and other marine migrants possess sensory equipment equivalent to that invented and used by human navigators.

The delay is not hard to understand. Specific open-ocean fish are not easy to find, tag, and track. Ships and crews are expensive, and captured animals are difficult to keep alive and work with in laboratories.

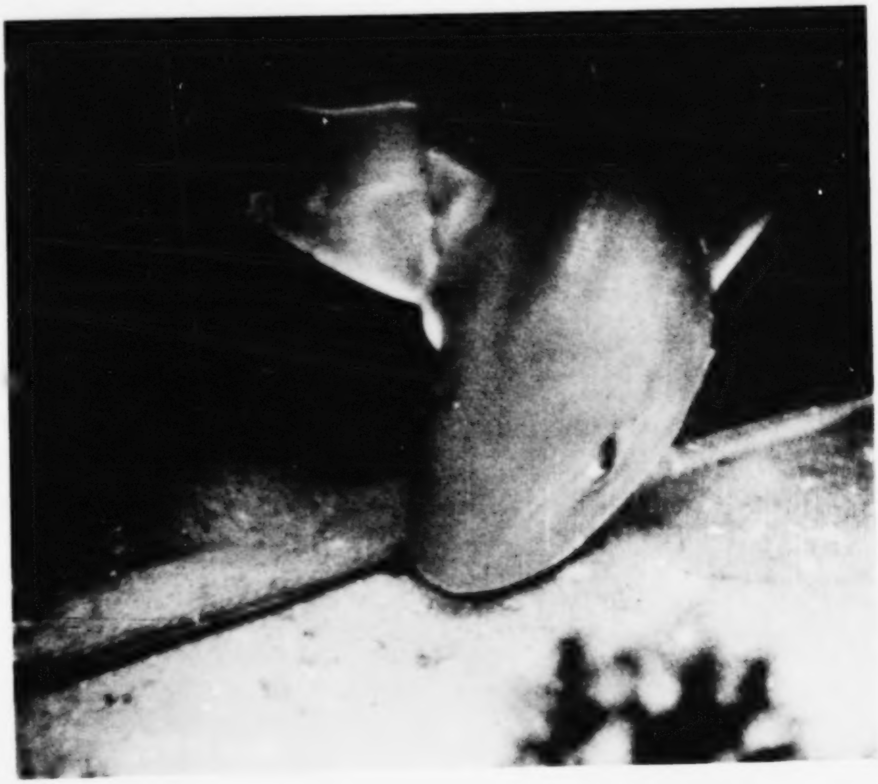
The incentive to study marine migrations did not emerge until fish became a critical source of protein for the world after World War II. But once concerted investigation began, the realization emerged that the seemingly trackless oceans contain a rich variety of path-finding cues. To exploit them, the seas' inhabitants have the same kinds of equipment that migrant or homing land animals and birds have, including solar and magnetic compasses, clocks, and polarizing lenses. In addition to being beyond the experience of humans, until the 1970s these sensory abilities in marine creatures were beyond the knowledge and imagination of most scientists.

Salmon are a case in point. It is commonly known that they make long and arduous migrations between the freshwater streams where they are born and spawn and the open sea. Before World War II, however, it was generally believed that they confined their salt-water ramblings to coastal waters, never straying far from their home streams.

This fit in with the theory, later proven by rigorous experiments, that a salmon finds its native stream and follows it to the spawning grounds by odor. Those experiments were done by Arthur D. Hasler, a University of Wisconsin biologist. Each of the streams in which salmon spawn and hatch, Hasler observes, acquires a unique fragrance from the plants and soils on the land it drains. This odor is imprinted on the nervous systems of the fish before they reach pan-size. The salmon seem to be able to pick up these unique scents and follow them back to their own spawning grounds.

Other than smell

Hasler conducted these experiments, and similar ones on trout, in the 1970s. But as early as the late 1940s Japanese and Canadian fishing crews were reporting large schools of salmon so far out in the North Pacific that they had to be well beyond the influence



Electric lure. A shark bites viciously at electrodes carrying currents that simulate the bioelectric fields of herring and other prey.

A. J. Kalmijn/Woods Hole Oceanographic Institution of even the most pungent river mouths. Later surveys and tagging programs revealed that Pacific salmon, feeding in mixed schools in the Gulf of Alaska, separate into groups of geographic origin and migrate thousands of kilometers to home spawning grounds in North American or Asian streams. North American pink, chinook, coho, chum, and sockeye salmon leave the Gulf, navigate through and around the Aleutian Islands, and swim 1,500 to 2,000 kilometers to the Yukon River. One record holder, an Asian pink tagged off Alaska, was recaptured 7,800 kilometers away in the sea between Japan and Korea. Tagging experiments in the Atlantic showed that salmon from United States and Canadian rivers routinely journey to Greenland and back.

Many observers proposed that such trips could be made by following ocean currents. But as more became known of the currents in the sea (See "Circulation of the Oceans," *Mosaic*, Volume 4, Number 3, and "Ocean Circulation: A Stirring Tale," *Mosaic*, Volume 10, Number 5), it became clear that these are far from the reliable, unidirectional highways they would have to be to become navigational highways. Further, swimming through a moving mass of water can provide little reliable information about speed and direction.

One thing ocean travelers can do, however, is check the sun during the day. Human and other navigators use this reference to find direction and latitude, accommodating to apparent shifting of the sun's position in the sky.

Bees do it

Hasler had worked with the German biologist Karl von Frisch, whose work in the 1950s disclosed that honeybees, using the sun as a travel guide, can compensate for its apparent daily motions. Hasler decided to find out if fish could somehow do something similar.

Honeybees, of course, get a much better view of the sun than do fish. Wave motion breaks up the radiant globe into a confusing cluster of shimmering lights. Refraction distorts the solar altitude and compresses the horizon from 180 degrees to about 98 degrees. But fish that leap out of the water to snap up insects evidently compensate for distortions, so Hasler decided to see if fish could use the sun as a compass.

He simulated the experience of oceanic fish by working with white bass in Lake Mendota, adjacent to the University of Wisconsin campus in Madison. Bass, displaced from two small spawning grounds on the north shore of this large lake and released near the center, face a navigational problem comparable to that of a bluefin tuna finding its way from the open Atlantic to coastal feeding grounds off North Carolina.

Hasler found that most of the white bass migrating from midlake maintained a relatively constant compass direction, regardless of the time of day or the season, "evidently by using the sun as a reference.... They clearly compensate for the changing solar azimuth as it appears to move from east to west," says Hasler. Hasler then took fish he had trained at Madison (43 north latitude) to Solemar, Brazil (24 south latitude). Experiments there showed that they could be trained to adjust to other latitudes, says Hasler.

The cue a piscine navigator uses to make this adjustment, according to Hasler, is the rate of change in the sun's altitude. Given some time in a local area, he adds, a bass or salmon can learn to adjust its course to the angle between the sun and a selected direction. "Salmon," he says, "moving from the Gulf of Alaska to rivers in North America or Asia, must make continual adjustments of that kind as they travel."

Beyond the sun

"That an animal can accomplish innately something for which people require instruments, charts, and tables is quite incredible," Hasler comments. More incredible is the fact that Lake Mendota bass and ocean fish find their way even when the sun is not visible. Similarly, salmon not only maintain direction in the North Pacific when the sky is completely overcast, which it frequently is, but sailors and fishermen have observed them traveling at night. Large predators such as tuna, marlin, swordfish, and sharks can hold a steady course and speed for days without coming up to check the sun.

"It seems to me that a sun compass is something most fish possess," remarks Hasler, "but they must receive and use a variety of other cues." These may include polarized light, changes in the earth's magnetic field, and natural electric currents.

The use of electric currents and fields was suggested by experiments Hasler did in the North Pacific in the late 1960s and early 1970s. He followed 43 sockeyes fitted with sonic transmitters and displaced 70 kilometers south of their destination, the Fraser River in British Columbia. "The most active individuals moved generally north," he says, "keeping away from the land and following the axes of near-shore currents with a bias toward their destination. Some sun-compass orientation must have been involved," he says, "since the salmon maintained northward progress despite reversing tidal currents. However, movement along current axes, night and day, under clear and overcast skies, suggests that fish also get directional information from currents." Since the clue could not have been the movement of the water because of the absence of fixed references, other clues had to be present.

Electrifying currents

When salt water, which conducts electricity, moves through the earth's magnetic field an electric current is induced in the water. Marine animals able to detect such an electric field could find their way around the oceans day or night, overcast or not, deeply submerged or near the surface. But is this generator effect strong enough to be useful? And do fish have the ability to use it?

To answer the first question, James D. McCleave of the University of Maine at Orono combined data on the conductivity of Atlantic Ocean water and the velocity of the Gulf Stream as it eddies north along the East Coast of the United States. According to his calculations, measurements of 0.005 microampere per square centimeter are common.

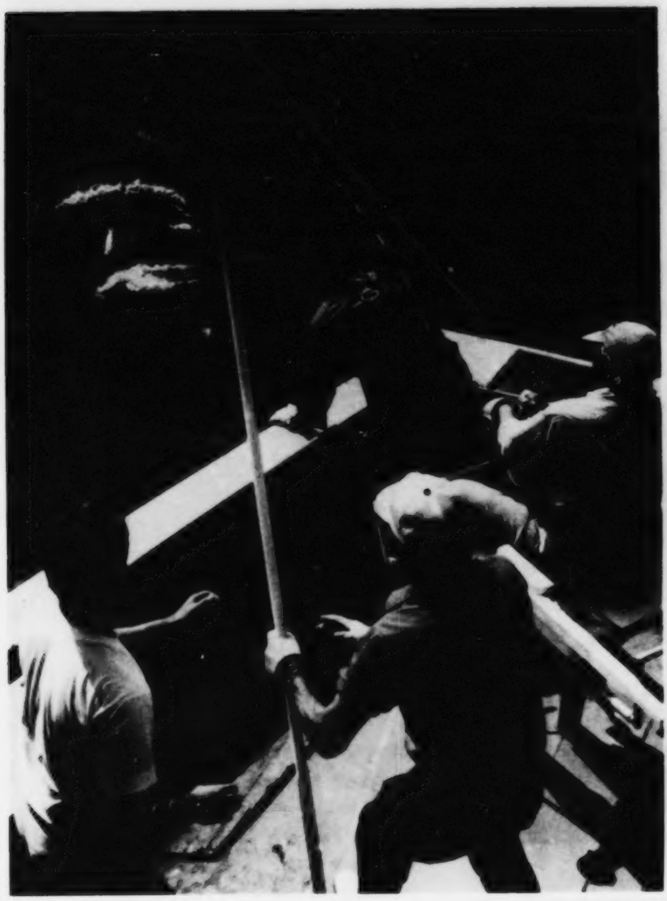
But can fish sense such currents? McCleave and Sentiel A. Rommel tested the reactions of Atlantic salmon and American eels to find out. They concluded that current densities as low as 0.0015 microampere per square centimeter, less than a third of the value commonly found in the ocean, can slow a salmon's heart beat, and density values not much more than that (0.00167) do the same to an eel. The researchers also determined that electrical stimuli "on the order of that probably generated by faster ocean currents" can make young eels veer to the left or right of the direction in which they would swim in the absence of such fields.

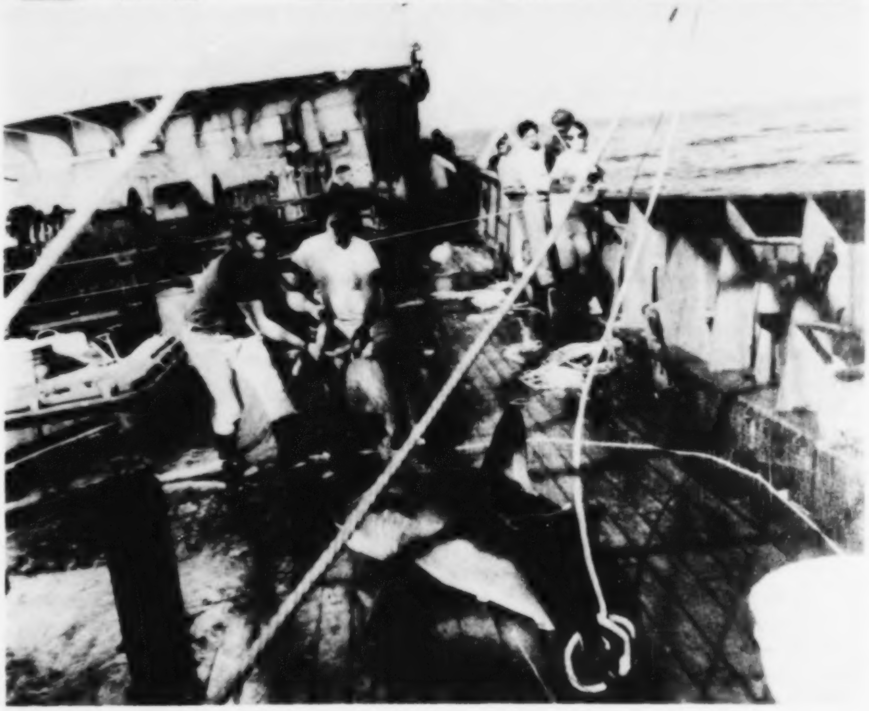
American and European eels migrate between freshwater feeding areas and ocean spawning grounds east of the Bahamas. McCleave believes that the adults and the newborns going to opposite sides of the Atlantic from the Bahamas may use the sun for navigation, but switch to electric cues when the sun is not visible. "The electric fields they would encounter would be perpendicular to the direction of current flow," he points out. "An eel would experience a rapid change in voltage when it turned its body from side to side as it swam."

This idea has not won wide acceptance; others who have tried to repeat McCleave's experiments have not been successful. Also, close inspection of dissected Atlantic salmon

---

Cromie is executive director of the Council for the Advancement of Science Writing.





Tagging. Researchers catch and tag marine species in order to track their migrations. Clockwise from top left: A bluefin tuna is caught off the coast of Quebec; James McCleave fits an eel with an ultrasonic transmitter; a transmitter is inserted into the stomach of a salmon; biologists cautiously prepare to tag a shark.

Barry M. Fitzgerald/International Game Fish Association; R. C. Kleckner/University of Maine; A. D. Hasler/University of Wisconsin; Francis G. Carey/Woods Hole Oceanographic Institution

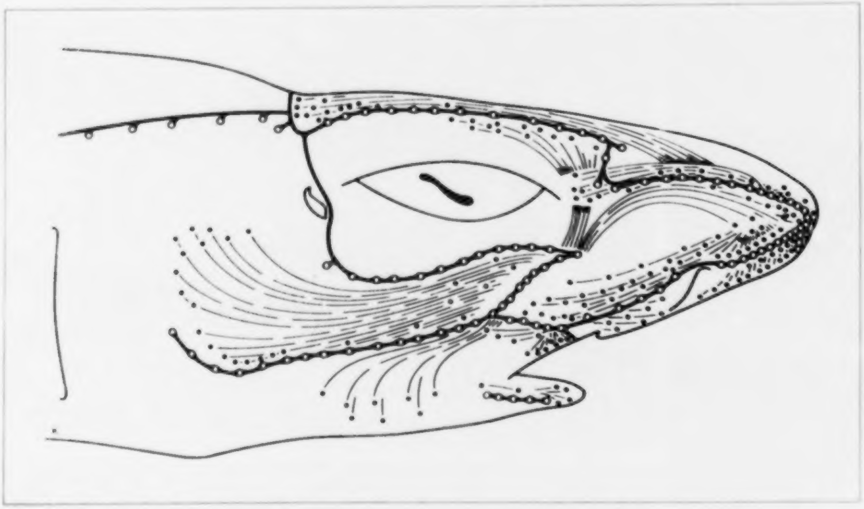
---

and eels has failed to turn up organs for reception of the vanishingly small geocurrents.

Electric systems are not unknown in the aquatic world, however. Certain freshwater catfish, electric eels, and some rays have electric organs that detect or stun prey and in some species are used for communication. None of these fish make long migrations across open water. But sharks travel great distances in the ocean, for reasons known only to themselves. And the electric sensors they possess may help them to get around.

Subtle sensors

The heads of sharks, as well as the capelike fins of rays, are peppered with pores



Receptors. The ampullae of Lorenzini (solid dots) found among lateral-line canals (heavy-black lines) and skin pores (open circles) of sharks are responsible for the animals' behavioral reactions to weak electric fields.

A. J. Kalmijn/Woods Hole Oceanographic Institution

leading into jelly-filled canals that end in swellings, or ampules, covered with nerve fibers. The functions of these organs, called ampullae of Lorenzini, were a mystery until the late 1950s work of Adrianus J. Kalmijn, a Scripps Institution of Oceanography biologist with a penchant for physics and electrical engineering.

Kalmijn first had to overcome the difficulties of generating and measuring electric fields on the order of fractions of microvolts per centimeter, smaller than the electric background in most laboratories. Once this was accomplished, he discovered that sharks and rays employ the ampullary system to sense such bioelectric fields, which are emitted by the activities of other creatures.

Rays, for instance, picked up the electric emanations of flounder buried in sea-bottom sand. Sharks vigorously attacked electrodes producing voltages simulating the bioelectric field of herring in a Massachusetts bay.

Kalmijn's investigations, conducted at Woods Hole Oceanographic Institution, demonstrated that sharks and rays are tickled by voltage gradients as small as a single volt in 2,000 kilometers. Fields generated by ocean circulation are 10 to 100 times as strong, so the fish would have no trouble detecting them. Nature therefore offers a prominent orientation cue, Kalmijn decided next to determine "whether these animals have the brain faculty and biological incentive to use this cue."

Sharks are, to say the least, difficult to capture and work with in a laboratory. Kalmijn chose to work with the small stingray *Urolopus halleri*. He trained these hardy, alert bottom dwellers to move right or left with respect to the direction of an electric field. A correct choice earned a morsel of herring; a wrong turn brought a "light but strict" prodding with a plastic rod. Stingrays also sense the earth's magnetic field, so some of the tests were run with this cue artificially eliminated.

Whether the geomagnetic field was nulled or synthetically altered, stingrays could select the direction that earned them a meal. His experiments, asserts Kalmijn, "unequivocally demonstrate that stingrays sense the direction and polarity of a steady, direct-current field as weak as 0.005 microvolt per centimeter, and execute educated orientational responses. Because such fields normally occur in their natural habitat, it is more than likely that sharks and rays exploit them in their daily lives." Kalmijn realizes that "such a conclusion remains to be verified in experiments conducted under oceanic conditions."

Animal magnetism

Sharks and rays do not confine themselves to drifting passively with currents; they swim about constantly. Moving with respect to the water, they also induce electric fields in their own bodies. "The ampullary system lends itself equally well to the reception of these self-generated fields," Kalmijn declares. And since the induced gradients depend on the intensity and direction of the geomagnetic field, the animals possess the equivalent of a built-in magnetic compass.

In his efforts to demonstrate this, Kalmijn is training stingrays to orient with respect to the earth's magnetic field in the absence of electric signals generated by current flow. Once they are trained, the magnetic field is turned off and the animals are presented with electric fields simulating those they would induce in themselves by swimming. "The ray is trained to find magnetic east; it must feel a particular sensation or voltage gradient when it swims in that direction," Kalmijn explains. "I'm trying to approximate this sensation and use it to guide the ray to a food reward." Kalmijn is continuing at Scripps where, he says, "my first priority will be to finish these crucial experiments with fresh stingrays and a larger tank."

Dual input from its own movement and from water flow do not confuse a shark or ray, Kalmijn says. Internally generated fields inform the animal about the direction in which it is heading, he believes, while externally generated fields reveal the direction of water movement.

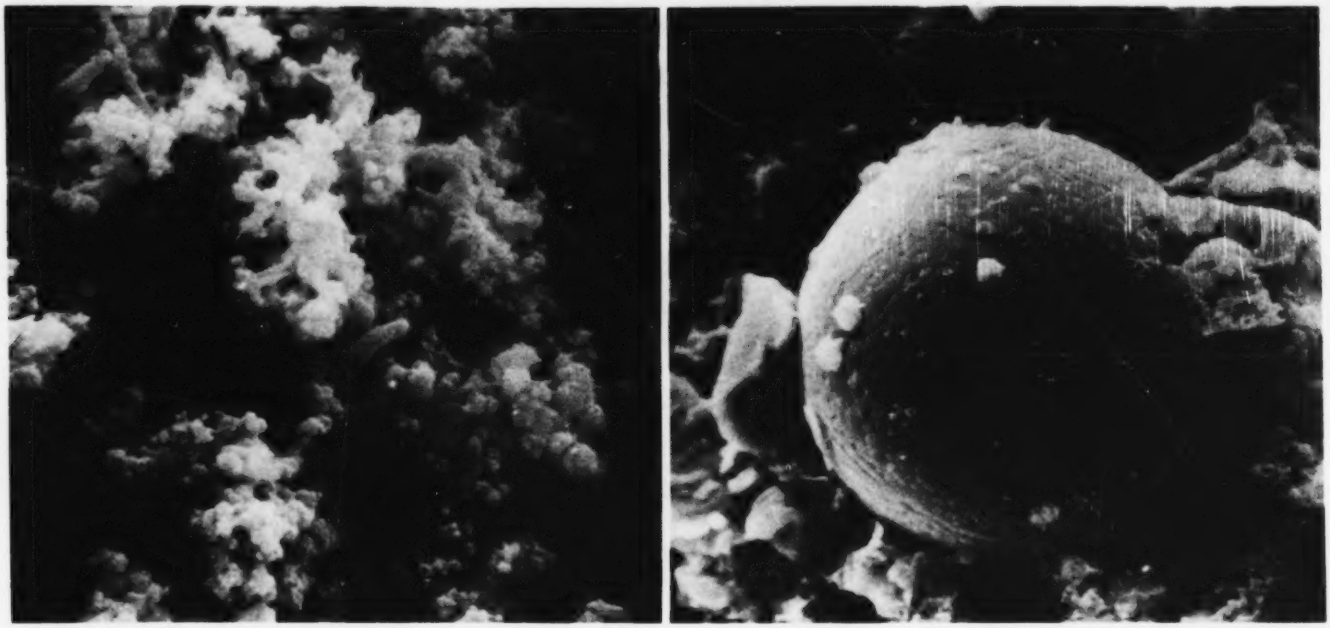
Nonetheless, questions remain—even if sharks and rays actually use such information for navigation. Most ocean travelers lack an ampullary system. How, then, do they find their way at night and in fog?

Magnetic senses

Frank G. Cary of Woods Hole Oceanographic Institution used sonic transmitters to track a swordfish that held a straight course over a 50-hour, 189-kilometer route. It approached the surface only at night and maintained a constant bearing while swimming eastward, across the Gulf Stream, despite a strong current set to the north.

In other observations, a bluefin tuna tagged off Cat Cay, Bahamas, was recaught only 50 days later off Bergen, Norway, more than 7,000 kilometers away. Juvenile yellowfin tuna travel a restless figure-eight path that takes them across the Pacific from Japan to California. Green turtles leave the coast of Brazil around December and swim day and night for 2,200 kilometers to reach tiny Ascension Island in the mid-Atlantic.

A magnetic sense to guide such journeys seems logical to a few researchers, but the scientific community generally has resisted the idea. "Before the 1970s," comments one biologist, "you could not talk about the influence of the earth's magnetic field on the behavior of animals without getting into trouble with your peers." Then a series of compelling experiments proved that some bacteria, honeybees, migratory birds such as European robins and warblers, and homing pigeons carry magnetic material in their bodies and use the earth's magnetic field as a travel reference. Homing pigeons, for ex-



Magnetic material. Particles ranging from 0.05 to 0.1 micron in size, from the skull of yellowfin tuna, may enable tuna to use the earth's magnetic field to determine direction. The function of the 5- to 50-micron diameter balls of magnetite (right) found in a number of marine species is a mystery.

J. L. Kirschvink (California Institute of Technology)

ample, get home under a heavy overcast, but not when they carry head coils that distort the geomagnetic field. (See "Animals Find Their Way," *Mosaic*, Volume 10, Number 3.)

In 1975, Richard P. Blakemore of the University of New Hampshire discovered that certain marine and freshwater bacteria respond like compass needles. They contain chains of magnetite or lodestone a few hundred angstroms in size, which lie in a line with a north- or south-seeking end. These lodestone lines orient to the inclination of the earth's magnetic field, the steepness of which increases with latitude.

Experiments conducted by Blakemore, Kalmijn, Joseph Kirschvink of Caltech and Richard B. Frankel of MIT have shown that, in the northern hemisphere, these bacteria move downward along magnetic field lines as they flick the flagella that propel them northward through the water. Their southern-hemisphere relatives, however, possess south-seeking compass needles that also guide them in the direction of the sea- or swamp-bottom. In either hemisphere, this orientation steers them away from what for them are toxic concentrations of oxygen near the surface and into the oxygen-poor, food-rich sediments where they thrive. In aquatic sediments near the equator, where field inclination is zero, north-seeking and south-seeking bacteria occur in roughly equal number.

The finding of this mechanism in bacteria intensified the search for magnetic receptors in large marine animals. Success was aided by the development of superconducting magnetometers sensitive enough to detect changes of ten parts in a trillion in a magnetic field. Such instruments have registered the presence of magnetic material in the heads of yellowfin and other tuna, blue marlin, green turtles, Pacific dolphins, one species of whale, and in human tissues. "Sharks also have considerable amounts of magnetite in and around their heads," notes Kirschvink, "opening the possibility that they possess separate sensory systems for detecting electric and geomagnetic fields."

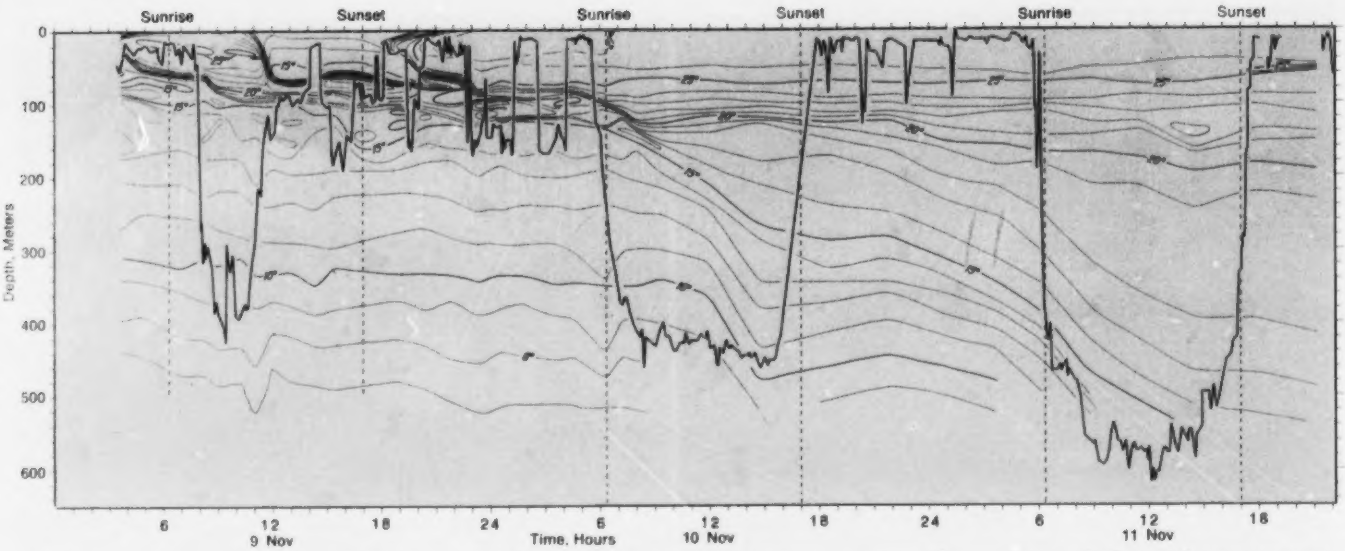
A head for directions

In their search for magnetic sensors, John Zoeger of Los Angeles Harbor College, along with J. Robert Dunn and Michael Fuller of the University of California at Santa Barbara, found strongly magnetized tissue in the heads of the common Pacific dolphin *Delphinus delphis*. Dissection revealed disklike particles on which scanning electron micrographs showed stalklike objects surrounded by ultrathin fibers. "The fibers may be a nerve net, and the stalklike objects could be part of a magnetoreceptor," Zoeger speculates. If this turns out to be the case, it could provide a clue to the puzzle of how whales find their way between feeding and breeding grounds thousands of kilometers apart.

Evidence for a magnetic guidance system is even stronger in tuna, marlin, and green turtles. Researchers using glass and plastic scalpels have dissected skulls of these tireless travelers and found millions of angstrom-sized grains of magnetite, as well as dozens of large spheroids of this naturally occurring magnetic mineral. The spheroids are between 10 and 50 microns in size and look like the tiny iron balls that originate from the breakup and melting of meteorites, says Kirschvink. "They are too large to function as compass needles; we can only speculate on their function, but magnetite is an excellent conductor, and the large spheres might be involved in a system that detects geoelectric fields," he suggests.

The large number of tiny particles might account for the magnetic acumen found in yellowfins in laboratory experiments designed by Kirschvink, along with Andrew E. Dizon and Michael M. Walker of the National Marine Fisheries Service. They found that tuna can easily be trained to discriminate between the normal geomagnetic field and a slightly stronger, artificial field.

From the number and size of magnets in the animals' skulls, and from the animals' laboratory performance, Kirschvink and his colleagues conclude that tuna, marlin, and green turtles possess more than enough magnetite to account for their direction-finding behavior. Two separate magnetic senses seem to be present, Kirschvink says. One senses direction; the other responds to small changes



Depth trace. A swordfish swimming west from North Carolina held a relatively straight course for 50 hours, but the depth record shows a dramatic change in level from sunset to sunrise.

Francis G. Carey/Woods Hole Oceanographic Institution

in the intensity of strength of the geomagnetic field. The latter apparently allows animals to detect small anomalies in the earth's magnetic field that they might use in the same way that humans use maps, Kirschvink proposes.

Tectonic street maps

The ocean bottom is a mosaic of mounds, ridges, faults, and rock boundaries which, Kirschvink says, could serve as magnetic seamarks. A worldwide system of midocean ridges and rifts defines areas where molten rock wells up from below to form new sections of the earth's crust. These sections generally consist of parallel, north-south avenues of rock that vary, from the ridges to the edges of the continents, in age and magnetization. Magnetization increases toward the ridge system and sharp anomalies occur at boundaries between differently magnetized sections. These anomalies, Kirschvink notes, can produce fields 10 to 100 times above the sensitivity threshold of pigeons and bees.

With sensitivity to these fields and the faculty to use it, fish might get information from the ocean floor comparable to the grid of streets and avenues on a map. Changes in dip could signal north-south progress past the equivalent of streets of latitude, while the anomalies mark off what could be avenues as a tuna, marlin, or turtle proceeds east and west. A cartoonist might depict one tuna telling another: "Hawaii is the fifth anomaly on the right; you can't miss it."

Such a scenario is largely conjecture, even though evidence suggests that homing pigeons and migratory birds use anomalies in this way. Definitive experiments still need to be done. Electronic devices must be fitted to tuna as they have been to birds so that scientists can shift magnetic environments at will. "This would allow us to determine if anomalies of the kind that exist on the ocean floor can change animal behavior," says Kirschvink.

There appear to be features on the ocean bottom that act as fences or road signs to migrating fish. "Albacore tuna swimming along the California coast," says Kirschvink's colleague Andrew Dizon, "seem to avoid crossing an escarpment running west from Cape Mendicino." The top of the ridge is 1,800 meters below the surface, far below the tunas' normal traveling depth, but it is marked by a dramatic magnetic anomaly.

A speculative "how"

Dizon is attempting to match such anomalous features with known concentrations of fish and other marine animals in the Pacific. Mike Walker, working at the University of Hawaii, is seeking to pin down tuna responses to magnetic changes of the magnitudes they would encounter along their normal travel routes.

At the same laboratory in Hawaii, Anjanette Perry is trying to learn if juvenile green turtles have a magnetic sense. Although they possess magnetic material similar to that of tuna, sea turtles do not respond to training as well. Perry suspects the difference indicates a lower level of trainability rather than a lack of ability to respond to magnetic cues. She won't concede that turtles may just not be as bright as tuna.

Generalizations, Dizon says, must await successes across species lines. "We have not made the tuna or turtle connection yet. We know what kind of magnetic material they carry and how they respond in a laboratory situation, but we don't know how they actually sense the geomagnetic field or use it for orientation."

How, for instance, are weak magnetic impulses coupled to the nervous system and integrated with other information? Numerous hypotheses exist, and Kirschvink champions his own.

Magnetite possesses the lowest resistivity of any known biological material, he explains. A 500-angstrom cube of it could depolarize a nerve membrane by shorting across it. Multiply such effects, he speculates, and "a few hundred particles would be sufficient for a built-in compass; millions of them would make possible the high resolution needed to detect anomalies on the ocean bottom."

Carrying such receptors and a sun compass, a piscine migrant might calibrate itself to a particular feeding or breeding site by combining observations of the rate of change in the sun's altitude, the earth's magnetic dip, and local magnetic anomalies. It would not be unprecedented; evidence exists, Kirschvink points out, that homing pigeons do this. To get from one familiar place to another, however, the traveler needs to find the correct direction in which to swim. At the start of a journey it might determine this direction by a solar or magnetic compass, or it might take successive map readings and work its way to a destination by a "you're getting warmer" scheme. While examples of beeline tracks exist, sea-animal navigators also meander widely, as if they were testing first one direction then another.

Salmon magnetism

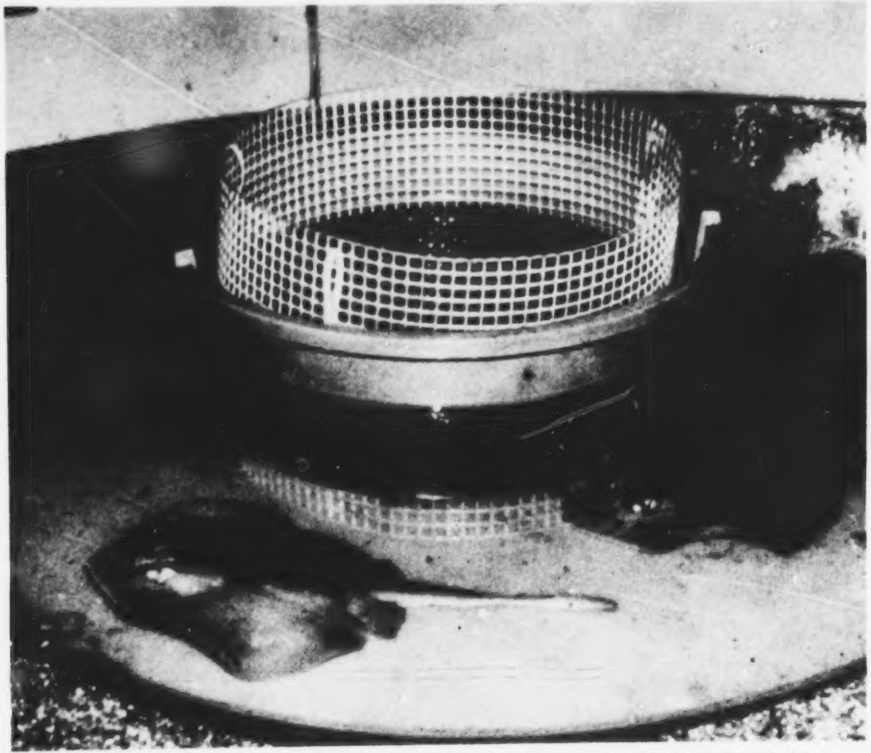
Do salmon get home this way? Magnetite particles have been found in their bodies, but researchers dismiss them as environmental contaminants. Nevertheless, experiments by Thomas P. Quinn of the Pacific Biological Station in British Columbia, working with Ernest I. Brannon and Ronald T. Merrill of the University of Washington at Seattle, indicate that sockeye salmon fry possess an inborn magnetic compass—even if it hasn't shown up in adults.

The fry sensitivity seems to be to the horizontal component of the earth's field, but experimental results "do not preclude use of the vertical component by adult salmon," Brannon notes. Tests at night and in tanks covered with black plastic lead the researchers to conclude that this magnetic sense guides fry on their nocturnal travels. During the day, orientation seems to be guided principally by celestial cues. "Fry use the magnetic field only under total overcast," Quinn says. "Our observations indicate that if a conflict occurs, they will believe visual rather than magnetic information."

The celestial preference apparently holds even when the sun is not visible. Sockeye captured while traveling to the ocean for the first time and placed in tanks have ignored artificial changes in the magnetic field and held to their courses on partly overcast days. Work by Kees Groot and Peter Dill of the Pacific Biological Station indicates that they may be able to do this by orienting to the patterns of polarized sunlight. This ability, alien to human experience but common in many species, is linked to the fact that the polarization of skylight is most intense in directions toward and away from the sun, least intense at angles to the sun of 90 degrees, and of intermediate intensity at other angles. Honeybees use this celestial cue, as do shrimp and other creatures that make short, daily migrations to and from feeding grounds.

Once sockeye reach the North Pacific, they encounter persistent cloud cover and may switch from a celestial to a magnetic preference, Quinn and Brannon speculate. "In the ocean, a population of salmon spreads out over thousands of square kilometers to feed," Quinn points out. Since they don't all come home from the same direction, a means of determining direction alone is not enough; they need an orientation sense to tell them where they are in relation to where they want to go. They should be able to use the same signposts as postulated for tuna and turtles. Quinn proposes that salmon follow a map grid based on dip (latitude) and the angle between magnetic north and true north (longitude) as determined from the sun's position. Either system requires sensitive magnetoreceptors, however, which have not been found in adult salmon.

Quinn and his colleagues discovered that salmon fry, fresh out of their gravel nests, use their magnetic sense preferentially on their first night of travel. The scientists also discovered that the presence of the moon enhances this orientation ability. "Whether the moon serves as a visual cue or somehow stimulates an innate magnetic perference remains a mystery," Quinn says. "The only things we know for sure are that the directional preference becomes stronger in moonlight and that this tendency to swim in a certain direction has survival value." It guides the fry to a nursery lake and aids their dispersal to feed and avoid predators. (Unlike other Pacific salmon, a sockeye spends its first year in a lake.)



Rays in a tank. Stingrays tested and trained at Woods Hole Oceanographic Institution circle in a tank surrounded by electric coils. The coils null the earth's magnetic field and generate artificial magnetic and electric fields.

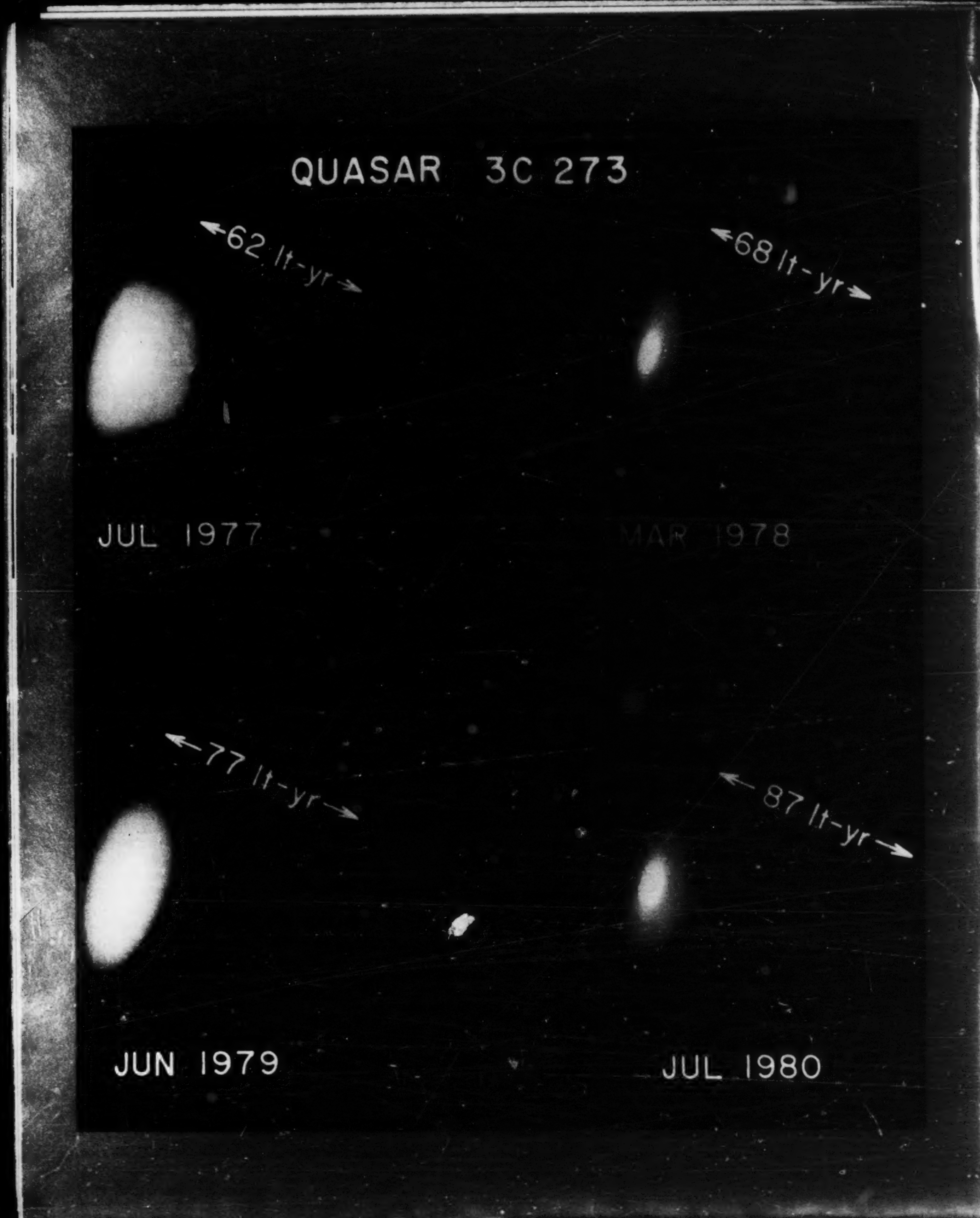
A. J. Kalmijn/Woods Hole Oceanographic Institution

Salmon thus use a complex, and so far bewildering, diversity of orientational cues. Their preferences change from magnetism at birth to celestial cues in adolescence and perhaps magnetism again as adults.

The diversity is not unique to salmon. Most marine migrants and travelers probably take advantage of different guideposts under varying conditions. Tuna, marlin, swordfish, and big sharks will often travel along a subsurface temperature boundary, or thermocline—a steep gradient between warmer, wind-driven upper water and colder, deeper, slower-moving water. Eels seem to use the boundary between two water masses of different temperature, salinity, and odor to identify their deep-ocean spawning grounds. When swimming or drifting near the surface, ocean travelers also might check wave patterns to determine the direction of currents driven by prevailing winds.

After more than 30 years of research, however, scientists still don't know how a fish finds its way home, and the mystery is increasing in complexity. Nonetheless, some long-standing prejudices are crumbling, and a deeper understanding is being gained of the sensory worlds of other animals and of the limitations of our own. •

The National Science Foundation contributes to the support of research discussed in this article principally through its Psychobiology and Oceanic Biology Programs.


QUASAR 3C 273
←62 lt-yr→
←68 lt-yr→
JUL 1977
MAR 1978
←77 lt-yr→
←87 lt-yr→
JUN 1979
JUL 1980

FASTER THAN THE SPEED OF LIGHT?

by Edward Edelson

The appearance of superluminal motion observed in some quasars and galaxies must be an illusion. If not, the laws of physics will have to be rewritten.

In October 1970, radio astronomer Irwin I. Shapiro of the Massachusetts Institute of Technology began a confirmatory test of general relativity. He was using what was then a recently developed observing technique called very long baseline interferometry. With radio telescopes 4,800 kilometers apart—at the Goldstone Deep Space Station in California and the Haystack Observatory in Massachusetts—Shapiro and his team set out to measure the deflection by the gravitational field of the sun of radio waves emitted by quasar 3C 279 in the constellation Virgo.

For several days, they recorded the radio emissions of 3C 279 as the sun passed between it and the earth. As a control, they made similar observations of another quasar in Virgo, 3C 273, whose emissions were not affected by the sun.

They found something unexpected. Interferometry, which in effect tied the two telescopes together to create a single antenna as big across as the thousands of kilometers between them, enabled the astronomers to view the structures of the two quasars in unprecedented detail. For the first time the compact cores of quasars could be seen to consist of several separate components.

Early in 1971, Shapiro mentioned his observations to Marshall H. Cohen, then at Caltech and one of the developers of very long baseline interferometry. Both Shapiro and Cohen soon applied the technique to additional observations of the two quasars, with even more unexpected results.

Shapiro was the first to publish his findings. In a brief note added to a report of his first observations, he wrote, "The separation of the components of 3C 279 appears to have increased by about 10 percent in four months." Prosaic as that might sound, it was a startling statement. Given the generally accepted idea of the size of the universe and the distance of the quasars from earth, it meant that the components of 3C 279 were moving apart at about 10 times the speed of light.

By common agreement, such rapid motion is impossible. It is basic to relativity that nothing moves faster than the speed of light.

One possible explanation that occurred to some astronomers was that the apparent superluminal expansion, a term commonly used if not coined by Roger D. Blandford of the California Institute of Technology, was an artifact, the result of faulty observations. That thought has since been abandoned. More than a decade of observations has amply confirmed the reality of apparent superluminal expansion. The phenomenon now has been observed in five quasars and one galaxy. If any conclusive demonstration of the cleanness of the observations were needed, it came in April 1981 when a large group of Caltech astronomers reported that after extensive observations they had produced a series of detailed maps of 3C 273. The maps provided "the first direct and unambiguous evidence of superluminal expansion in any radio source," they said. What is being seen—whatever it is—is real.

Most scientists thinking about superluminal expansion these days are seeking an explanation of what is seen that would not require velocities beyond the speed of light. "Nobody wants to violate that law of physics," says astronomer Geoffrey Burbidge, now director of the Kitt Peak National Observatory. Unfortunately, however, each of the two alternative explanations most widely considered feasible has what is probably a fatal flaw.

The Hubble wall

The first explanation is advocated by a minority of astronomers that includes Burbidge and Halton Arp of the Mount Wilson and Las Campanas Observatories in California. It explains the illusion of superluminal expansion by bringing the objects displaying it much closer to earth than their apparent astronomical distances. If the objects showing superluminal expansion are, say, only a twentieth as far away as they seem, then the speed with which their components are moving apart would be only a twentieth as great as had been calculated.

That explanation, as neat as it is, is at least as difficult for most astronomers to accept as is superluminal motion itself. To embrace it, they would have to accept the idea that, for unknown reasons, the uniform standard used to measure astronomical distances—Edwin P. Hubble's red-shift-based constant—does not apply to some selected quasars and galaxies. And exceptions don't prove rules: they destroy them.

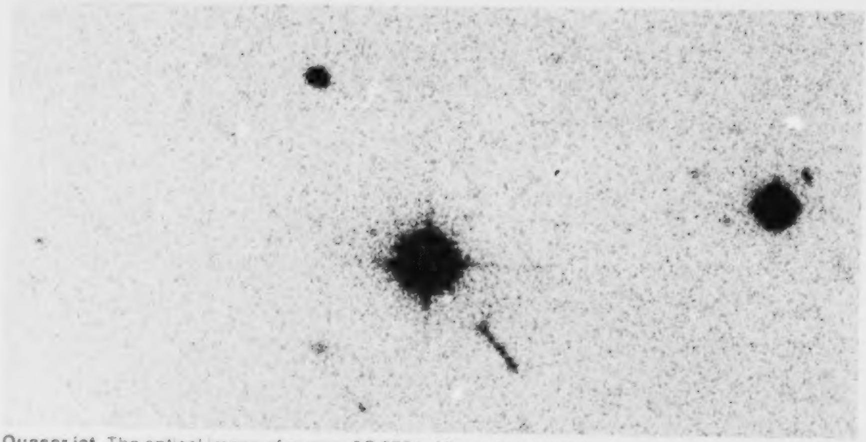
The universe is known to be expanding rather like a balloon that is being inflated steadily. Hubble showed half a century ago that astronomical objects move apart in that expanding universe at velocities that are in direct proportion to the distance between them. The most distant objects move apart with the greatest velocities.

If an object is moving away from an observer, its light waves are perceived as stretched, which shifts their light toward the longer, or red, end of the spectrum of

---

Superluminal motion. Radio maps of quasar 3C 273, taken over a four-year period, show a gaseous knot moving away from it at almost ten times the speed of light.

T. J. Pearson, S. C. Unwin, M. H. Cohen, et al./Nature



Quasar jet. The optical image of quasar 3C 273 in Virgo shows a jet that may be part of the explanation of the object's bizarre behavior.

Halton Arp

visible light. Astronomers thus can determine an object's distance from earth by measuring its red shift and applying Hubble's constant, producing relationships of velocity, distance, and, by inference age, that are among the best respected relationships in science.

Quasars are distinguished by red shifts so large that the Hubble constant places them among the most distant objects known. But Halton Arp proposes that the red shifts of some quasars are not what they seem to be, are not embraced by Hubble's rule. Despite their red shifts, Arp says, these quasars are relatively close to the earth; another—and unknown—mechanism must be responsible for their red shifts. That mechanism can be explained only by invoking some new and revolutionary concepts. It would preserve one apparent universal by violating another.

Relativistic cannonballs

Rather than joining Arp in requiring a revision of basic physical laws, the majority of astronomers prefer to explain superluminal expansion by one version or another of what Martin J. Rees, director of the Institute of Astronomy in Cambridge, England, has called the relativistic cannonball.

Rees holds that superluminal expansion appears to occur when a distant, or cosmological, object expels a jet of matter at relativistic speeds—velocities approaching the speed of light. If such a jet were headed almost directly toward an observer, Rees says, the apparent difference in arrived time of signals from the moving and the stationary source would make them seem to be separating at a rate faster than the speed of light.

As attractive as it appears, that proposal has at least two difficulties. First, no astronomical object emitting a jet that moves at nearly the speed of light has ever been directly and unambiguously observed; second, if one were, it would be a more-than-remarkable coincidence to find that the jets from all superluminal sources known were pointed with uncanny accuracy almost directly toward earth. The likelihood of that happening, says Kenneth I. Kellermann of the National Radio Astronomy Observatory, has a probability of only a few percent. Nonetheless, even with its improbabilities, adds Kellermann, most astronomers agree that the relativistic cannonball is a likelier explanation than a noncosmological nature for some quasars and galaxies, as Arp and Burbidge offer.

The cannonball proposition was conceived long before superluminal expansion was first observed. It came out of an effort to solve another problem posed to astrophysicists by quasars: the inexplicable source of phenomenal energy of these exotic objects.

Quasars were first identified optically in 1963 by Maarten Schmidt of Caltech. The belief that they emit huge amounts of energy is based on their apparent distance from earth, as measured by their red shifts. At their presumed distances, quasars would be invisible if they weren't emitting great amounts of visible energy. And after nearly 20 years, there still is no widely accepted model of how such intense energy might be produced. Most astronomers believe that the mechanism must involve black holes, stellar remnants that have gone into infinite collapse because of gravitational forces. (See "Stellar Ontogeny:...to Ashes," *Mosaic*, Volume 9, Number 3.)

Arp's explanation of quasar energy sources is the same as that for superluminal motion: that the red shifts are not the consequence of their distance. If the quasars were noncosmological—relatively near neighbors of earth, rather than being on the edge of the visible universe—their energy emission would be rather modest and easily explained.

To make his point, for 15 years Arp has been collecting examples of quasars with very large red shifts that are apparently closely associated with galaxies that have small red shifts. He has an impressive collection of sky photographs in which high-red-shift quasars and low-red-shift galaxies seem to be linked either by proximity or by bridges of matter. Arp says that the associations are real and that the quasars are alongside the galaxies despite the disparity in the red shifts. His critics—a majority of the astronomical community—say that the associations are not real. They say the quasars lie behind the galaxies, and that Arp's bridges only seem to link bodies that appear to lie close to each other because they lie along the same line, though at greatly varying distances.

Arp's argument suffered a setback recently because of observations of an object in the constellation Taurus called 3C 120, one of the six known superluminal sources. For several years, there was a debate about the nature of 3C 120—quasar or galaxy. The new observations (made by, among others, Geoffrey Burbidge's wife Margaret, who is at the University of California at San Diego), showed convincingly that 3C 120 is not only a galaxy but is one that is relatively nearby.

Even though 3C 120 is much closer than the quasars displaying superluminal expansion, it still is distant enough for its apparent internal motions to require an exotic explanation. The components of 3C 120 appear to be moving apart at about four times the speed of light, compared to about 10 times the speed of light for the quasars.

Geoffrey Burbidge is undaunted. "I can only say that I was not radical enough," he asserts. "I should have said that some galaxies also have noncosmological red shifts. You've got to start worrying about the red shifts of galaxies as well as those of quasars. That's a problem Arp's faced up to but I haven't."

The majority of astronomers, who find it difficult to accept Arp's contention that a completely nonstandard set of physical laws is working in some quasars, are even more reluctant to include galaxies in that unusual category. They say that, in proposing so many exceptions to the rules, Arp's theory comes to resemble Ptolemy's invocation of cycles and epicycles to explain the motion of the planets and the sun before the heliocentric nature of the solar system was accepted.

Edelson is former president of the National Association of Science Writers.

Relativistic jets

Several other lines of evidence are cited against the Arp theory. One is the fact that Martin Rees predicted the existence of apparent superluminal expansion in 1966, five years before the first observations were made. "We knew that some radio sources consisted of relativistic gas, and if you have something that expands at close to the speed of light, you can say that it will appear to expand superluminally," Rees notes.

Rees's assertions have been strengthened by observation of SS 433, an unusual star in the constellation Aquila, in the Milky Way galaxy. These observations in 1978 by Bruce Margon, now at the University of Washington, showed that some of the light emitted by SS 433 was red shifted and some was blue shifted. The blue shift is the red shift's obverse: evidence that an object is approaching rather than receding from an observer.

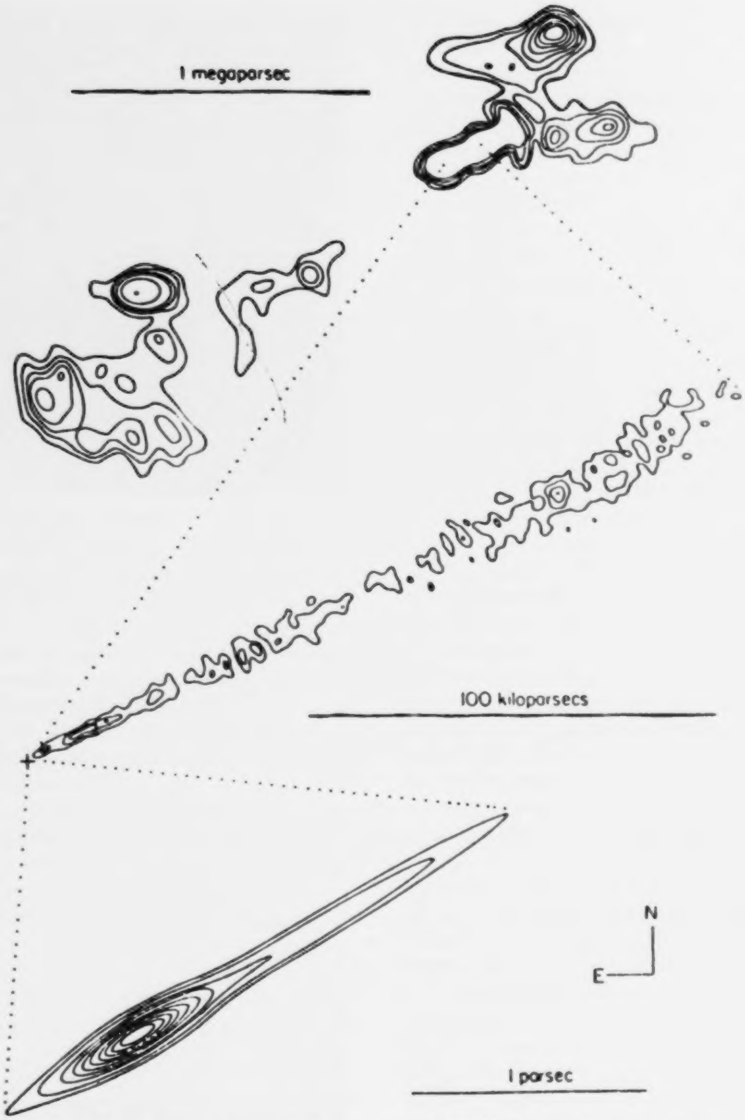
If its emission lines were to be taken at face value, SS 433 would have be be moving simultaneously toward and away from the earth. Nobody thinks this. A series of detailed observations, however, has shown that SS 433 appears as it does because it is ejecting two narrow jets of matter, one toward the earth and one away, each moving at about one-quarter of the speed of light.

Marshall Cohen says the emission of such fast-moving jets by an otherwise inconspicuous star is, by extension, a telling argument for the emission of even faster-moving jets by such larger and more energetic astronomical objects as quasars and galaxies. "It is comforting to know that such a thing could exist on a small scale," Cohen declares. "Otherwise there would [have to] be an enormous jump to relativistic jets [in quasars]. With SS 433, the extrapolation does rot have to be so big."

Several ways in which relativistic jets can be produced have been proposed. Most of them postulate the existence of black holes that are from a million to a hundred million times as massive as the sun.

Gas in the region of such a black hole would tend to form a disc because of gravitational effects. As Roger Blandford of Caltech explains, some of the gas would drift in toward the black hole and then be deflected in a jet. If there were a large enough amount of gas around the black hole, the innermost portion would pile up in a doughnut-shaped cloud that would funnel the jet into a tight stream. That kind of jet could have the relativistic speeds needed to give the appearance of superluminal expansion.

Alternatively, gas flowing into a black hole could be repelled in a tight jet by magnetic-field effects, Blandford says. A spinning black hole could behave like a giant flywheel, passing much of its energy into a relativistic jet.



Increasing resolution. Contour maps of elliptical galaxy NGC 6251 show the resolving ability of conventional radio interferometers (top two), compared to that of very long baseline interferometry. The VLBI map resolves the heart of the object's energy source.

Readhead, Cohen, Blandford/*Nature*

As in some short-answer tests, all of the above may be true because of the large variety of jets that have been observed. Some are emitted by stars, some by galaxies, some by quasars. "There could be many different mechanisms involved in the production of jets," Blandford notes.

The coincidence

One equivocal bit of evidence has come from observations of the most recently discovered superluminal source, the quasar found in the constellation Perseus and designated NRAO 140.

The X-ray detection of NRAO 140 was by the orbiting Einstein X-Ray Observatory. Alan P. Marscher, now of Boston University, and John J. Broderick of Virginia Polytechnic Institute (who subsequently found 140 to be superluminal) made very long baseline radio interferometry observations to determine the cause of the quasar's X-ray emissions—whether it was a known effect in which photons are nudged to X-ray frequencies by collisions with electrons, or some more exotic mechanism. They found that NRAO 140 emitted only a fraction of a percent of the X-ray energy that should be detected if conventional mechanisms were at work. Considering the weakness of the X-ray emissions, Broderick and Marscher concluded, the radio-emitting region must be moving at something like 97 percent of the speed of light.

But what about the element of coincidence, the still-unaccounted-for statistical improbability that all observed jets are pointed toward earth with sufficient precision to produce the effect of superluminal motion? One possible explanation, Kenneth Kellermann says, is that some focusing effect is taking place and all of the radio energy is beamed within a narrow cone along the direction of motion. Because of this beaming of the radiation, only those sources with motion aligned along the line of sight would appear bright.

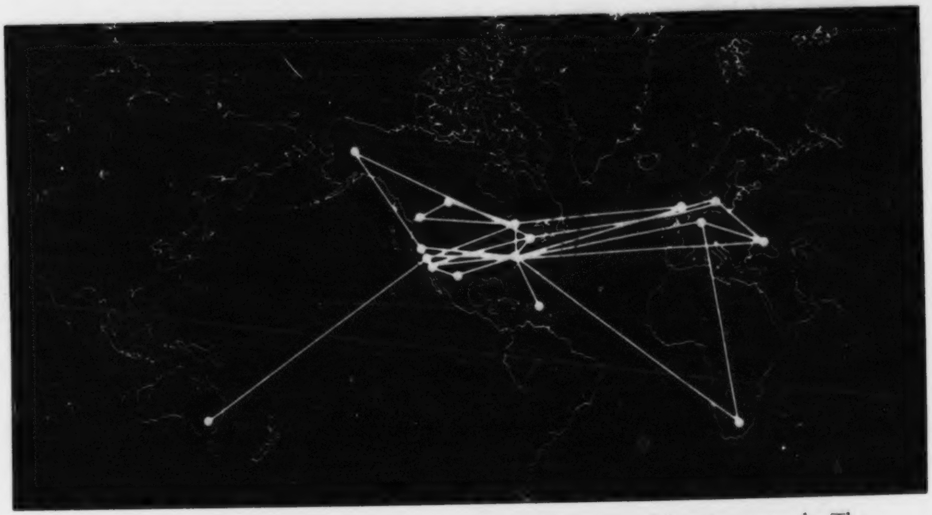
Since the investigations of superluminal motion have been confined to only the brightest radio sources, this explains why such an unusually large fraction of observed sources show the superluminal effect, says Kellermann. But superluminal motion is observed in the quasar 3C 273—the brightest quasar in the sky—at optical, infrared, and X-ray wavelengths as well as radio. And it is highly unlikely, he adds, that this unique object just happened to be pointed toward the earth.

Martin Rees has proposed that the radio sources actually emit radiation in a wide cone with a strong gradation of intensity. In this event, no matter what the angle of motion, part of the cone would be pointed earthward. By the same token, however, only a portion of the energy would be reaching the earth. The objections to this are that so large a cone, carrying so much energy, only intensifies the mystery of quasar energy magnitudes that has haunted astrophysicists for two decades.

Another possibility is that the superluminal illusion is produced by a gravitational lens, the apparent magnification of motion by the gravitational field of a cosmic body. Either way, there is a "strong consensus," Roger Blandford says, that the relativistic jet theory, in one form or another, turns out to be the correct explanation.

Truths, tools, and techniques

It is likely, as the studies continue, that more possibilities will emerge. Perhaps they will be as reasonable—or as unreasonable—as those now being considered to be the best. Probably the inconsistencies will become subtler and subtler until one is accepted as true.



That it will be an explanation constrained by the known velocity of light is considered probable. It is also probable that an answer will not emerge until the power of the available basic observational instruments—very long baseline interferometers—is much improved; to greater and greater extents, the improvement of scientific instrumentation and scientific understanding seem to go hand in hand. The explanation may also come from more sophisticated ways to evaluate the information that becomes available.

Radio astronomers first developed very long baseline interferometers to improve the resolving power of their telescopes. Now the pressure is reversed. The technique must grow to help explain what it has revealed.

The key to VLBI is precise synchronization of the separate recordings made at different telescopes. Early atomic clocks made it possible to synchronize clocks at radio observatories great distances apart with accuracies measured in microseconds. The effective diameter of the antenna then could be several thousand kilometers.

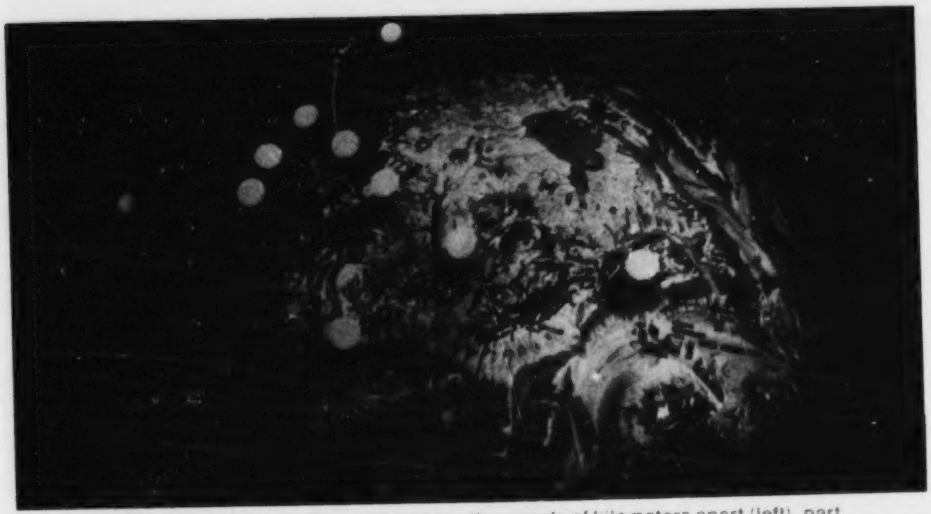
Early VLBI results were quite imperfect by today's standards. But atomic clocks have become much more accurate, and the technology has been improved over the past decade. With the first atomic clocks, recalls Timothy J. Pearson of Caltech, two telescopes could be synchronized to one part in 10^{13} for about two minutes. Now telescopes can be synchronized 100 times as precisely for 30 minutes or more. But more than improved matching of hardware seems to be needed.

Math answers

One nontechnological approach, a mathematical solution, has been found to compensate for the inevitable inaccuracies of even the most advanced instruments. Arriving radio waves from distant astronomical objects

Superluminal galaxy. A galaxy in the constellation Taurus, named 3C 120, has components that appear to be moving apart at four times the speed of light.

Halton Arp



Worldwide participation. Radio telescopes thousands of kilometers apart (left), part of the very long baseline interferometry network, can be linked together in various configurations to create, in essence, single instruments continents wide for the study of extremely compact radio sources. A proposed expanded VLBI array (right) will improve the study of radio sources at the shorter wavelengths and higher resolution.

NRAO

are affected by the interplanetary medium, by the earth's ionosphere, and by their passage through the atmosphere. Each has an impact on the received signal. "For a long time it looked as if you would never be able to make a proper image with very long baseline interferometry," says Anthony Readhead of Caltech. "The best you could do was make a model and see how well the data fit your model."

The goal was to synchronize the separate observations so closely that the data could be used to draw maps of the objects being observed. That the goal has been achieved is testified to by the detailed maps of 3C 273 by Pearson, Readhead, and their Caltech colleagues.

A part of the solution, developed early in the technology, was the use of more than two telescopes. In the 3C 273 observation, for example, both four and five radio telescopes were used in combination. The resulting tapes of radio observation were taken to Caltech, where up to five tapes can be played and synchronized simultaneously. But what happened next was even more important: the application of a new mathematical method developed by a team including Readhead and Peter Wilkinson of Jodrell Bank Observatory and Alan E. E. Rogers of the Haystack Observatory in Massachusetts.

The essence of the problem was to coordinate the phases of the radio signals, the time at which their peaks and troughs arrive at each of the telescopes. In a three-telescope array, there are inevitable phase errors among the instruments.

But the phase errors could be made to average each other out if the telescopes were taken a pair at a time. If atmospheric effects change the interference pattern of two of the three telescopes by a certain amount, the interference pattern of one of those paired with the third will be changed by an equal and opposite amount. The effect repeats as pairings of the telescope are varied among possible combinations. And as more telescopes are added to the array, the accuracy of the results increases. "Instead of a phase for each baseline, we take the sum of all baselines," Readhead explains. "As you go around the loop, the errors cancel out exactly."

Order from chaos

Yet another mathematical technique was developed to get true maps of superluminal sources from the radio telescope data. Despite strong views to the contrary, "The early results were almost made up in people's heads," says Pearson. And not unexpectedly the first attempts to draw detailed maps of the radio sources produced widely divergent results.

The change that made the difference seemed obvious, once it was discovered. The first mathematical treatments of the data had not excluded the possibility that parts of the sky could have an improbable something called negative brightness. It might have been called anything; it amounts to an index of wrongness somewhere in the model. "If the the sky could have negative brightness, an infinite number of solutions was possible," Readhead explains. Once it was specified that negative brightness was impossible, order emerged from chaos. The way it worked was this: The making of maps is an interactive process in which the astronomers first draw a map which invariably has some small areas of negative brightness. If the areas of negative brightness increase on the second try, the astronomers know they are heading in the wrong direction. As the negatively bright areas shrink, the map becomes more accurate.

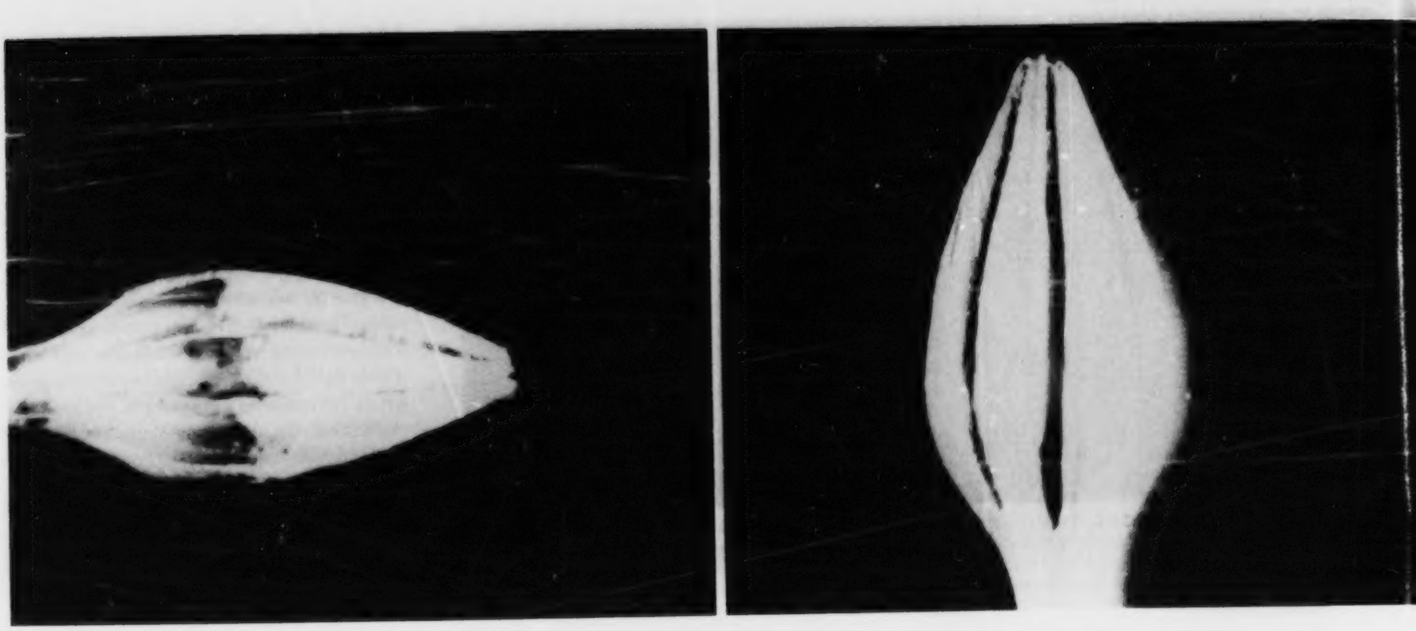
But even with these manipulations, the major problem with very long baseline interferometry now, the radio astronomers say, is that there isn't enough of it, especially when it comes to studying superluminal expansion. To detect such expansion, the same objects must be observed repeatedly over months or years. One reason for the small number of objects found to be displaying superluminal expansion is that observing time on very long baseline interferometer arrays is limited.

Readhead and Pearson hope to add to the number of sources. Since 1978, they have made initial very long baseline interferometry observations of more than 30 objects that had not been studied before. They will look at some of the quasars again at the end of this year to see if superluminal expansion can be detected in any of them.

Several proposals have been made for an expanded very long baseline interferometry array. The proposal being most seriously discussed is for a ten-telescope array, stretching from Massachusetts to Hawaii and from Alaska to Texas. One advantage of the proposed transcontinental radio telescope is that it would give astronomers better access to the shorter radio waves, "which is where the real action seems to be," Readhead says. "We're now using telescopes built for six-centimeter waves and pushing them to three centimeters. We would like to get down to one-centimeter or seven-millimeter waves."

In the longer run, astronomers would like to extend the baseline of interferometry even further by adding an orbiting radio telescope to the array. A number of potential scientific benefits might accrue from such enlarged arrays, Blandford says, and one of them could be an answer to still-unsolved riddles about quasars. "If the problem of the relativistic jets were solved, if we knew what was coming out and why, there would be a far stronger constraint on the design of the central machine," Blandford says. •

The National Science Foundation contributes to the support of research discussed in this article through programs in its Division of Astronomical Sciences.



To Land, to Seed, to Flower

by Boyce Rensberger

Plants had to conquer the earth's rocky terrain to set the world stage for the animal life that came along next.

It is early in the Silurian period, about 430 million years ago. The earth is poised for one of the great turning points in its history. The planet is already some four billion years old, and life has been evolving in its oceans for at least three-quarters of that time. The seas teem with plant and animal life. Simple algae may live in some moist regions, but as yet the land remains virtually barren.

The first life truly to conquer the rocky and lifeless dry land has to have been a plant. Plants are the only form of life able to capture and store the energy that feeds animals. Until plants went ashore, no animal could.

Once that start was made, however, in the next 230 million years—a mere 8 percent of the time that life has existed up to this point—all the major forms of land animals appeared, including mammals and birds. It was a veritable explosion of evolution. But none of it could have happened unless the plants went first.

As important as plant evolution was and is, relatively few scientists study it. One of those who do, Leo J. Hickey of the Smithsonian Institution in Washington, D.C., estimates that there are ten times as many zoologists as there are botanists and that only about one plant scientist in ten works on evolutionary questions.

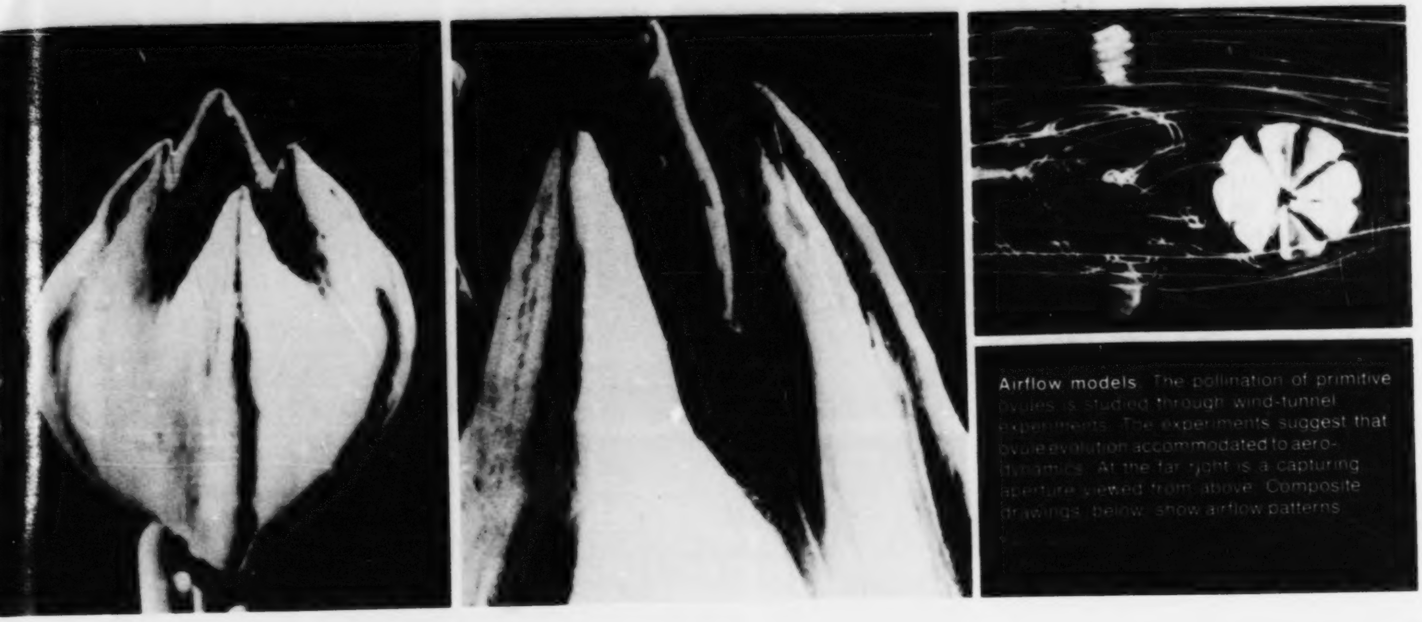
Hickey believes the low number of plant evolutionists is a result of the common attitude that plants just are not as interesting as animals, that they are a kind of second-class form of life. But even a cursory look at what is known about plant evolution—and at the tantalizing questions that remain unanswered—suggests that in many ways plants are among the strangest forms of life. At least from the typically zoocentric viewpoint, they are the strangest that one is ever likely to encounter.

Death-defying

For one thing, in some plant species the usual concepts of the individual organism or of inevitable death are meaningless. Through vegetative reproduction by underground stems or bulbs, a plant that sprouted from one seed may live for thousands of years. Some botanists think that patches of buffalo grass, a sod-forming species in the American plains, are at least 10,000 years old—far more ancient than the bristlecone pines. They probably sprouted from seeds at the end of the last ice age. Though portions of the patch have died, other areas—part of the same genetically unique individual—live on.

Sod grasses and such bulb-forming plants as daffodils die only as the result of outside forces. Disease or climatic change can kill them, but aging has no effect. Thus it is conceivable, according to G. Ledyard Stebbins, an evolutionary botanist at the University of California at Davis, that some individuals of these species live for millions of years.

A related difference between plants and animals is the fact that virtually all plants



Airflow models. The pollination of primitive ovules is studied through wind-tunnel experiments. The experiments suggest that ovule evolution accommodated to aerodynamics. At the far right is a capturing aperture viewed from above. Composite drawings, below, show airflow patterns.

are indeterminate in growth. This means they maintain clusters of embryonic cells, as in the growing tips, that are always capable of changing into roots, stems, leaves, or any other specialized tissues. The higher animals, by contrast, convert all cells into specialized tissues and lose the ability to grow or replace damaged areas. Senescence dooms the individual animal. Plants, with rare exception, never yield the embryonic option and thus are always capable of surviving severe assaults or even of regenerating the whole individual from a small fragment.

The land habit

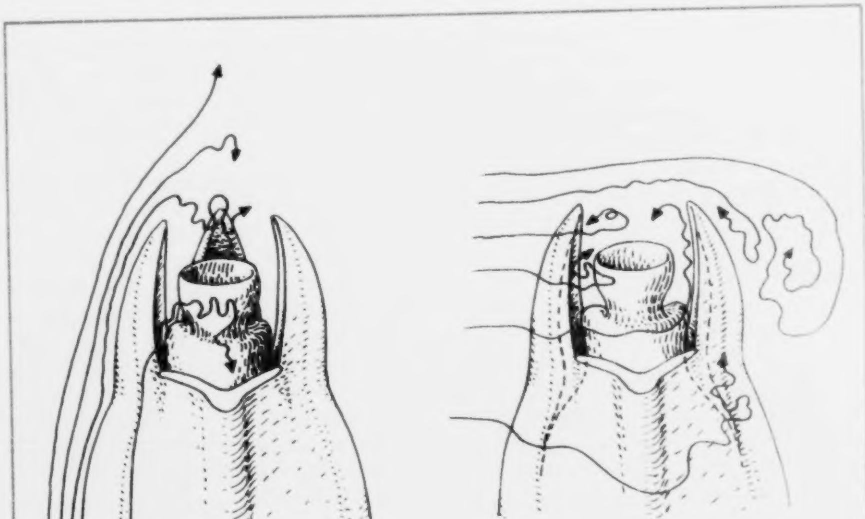
Until the Silurian, the only plants were algae. The simplest were one-celled forms with whiplike flagella for propulsion. They behaved like animals with chloroplasts, the photosynthesizing machinery that provided them with food. There were also motile colonial forms with head and tail ends. And there were the much larger algae, such as the seaweeds. None of them, however, was capable of living on land.

The land habit, as botanists call it, requires a host of adaptations unneeded in water. Indeed, much of the evolution that plants have undergone on land can be seen as a series of efforts to cope with the potential for desiccation and the lack of buoyancy in air.

Land plants need a means to absorb water and a system of cells able to conduct it to other parts of the organism. They also need some kind of waterproofing to keep their exposed parts from drying out. Algae, perpetually bathed in water, have no need for any of these adaptations; each cell absorbs its own water. Land plants need some internal rigidity if they are to rise much above the ground. Algae, buoyant in water, are otherwise limp. Land plants need some way of getting rid of the toxic-waste products of metabolism, while algae simply excrete them directly into the water. Land plants need some protection from the ultraviolet rays of the sun; algae have the shield of water.

"Getting on land was a major evolutionary hurdle and we really don't know exactly how or when it happened," says Karl J. Niklas of Cornell University. "But we do have some clues." Niklas specializes in the biochemistry and architecture of fossil plants.

The oldest known undisputed land plants date from the late Silurian, just over 395 million years ago. But Niklas has recently



found some interesting evidence among early Silurian plant fossils from Virginia (35 million years older still). It seems they had features that could have prepared them for life on land.

Fossil polymers

Plant fossils often contain trace amounts of biochemicals that resisted complete breakdown. In the early Silurian fossils, Niklas has identified compounds that resemble lignin. This tough substance lines the tubelike tracheids that conduct water up from the roots of modern vascular plants. Lignin is the stuff of wood: it gives plants the rigidity needed to keep a stem erect.

Lignin, Niklas says, is a polymer of certain plant metabolites which, if produced in high enough concentration, are toxic. He speculates that land plants polymerize these potentially toxic substances, yielding the relatively inert lignin. The lignin is then sequestered into specialized cells of the plant.

Such a metabolic adaptation to life out of water would have yielded an architectural and ecological advantage: the rigidity needed to grow taller. Niklas notes that these secondary effects would have had immediate survival value, because a taller plant could rise above the competition from the nonrigid, ground-hugging forms that may have covered the land at first. Height would also have been an advantage in dispersing dustlike reproductive spores from the top of the plant. Further, lignin and biochemically related substances, Niklas notes, are hard to digest and potentially toxic to animals. This would enable the early land plants to avoid being eaten by the animals for which they created a habitat.

There is no direct evidence that the early Silurian plants Niklas studied were living on land. The fossils lack morphological evidence of a water-conducting system. However, the presence of ligninlike compounds suggests that the plants, even if they were living in water, had a new metabolic ability that would later prove a mechanism for surviving on land.

Mosslike, liverwortlike

The traditional story of the botanical conquest of the land has suggested that the first plants to go ashore were multicellular green algae of the sort that form pond scum today. The larger and morphologically more complex red and brown algae, commonly known as seaweeds, are disqualified as ancestors because they are biochemically quite different from land plants.

Recently, however, G. Ledyard Stebbins has suggested that land plants may have evolved from something like the single-celled green algae that live in moist soils today. Some of their internal structures more closely resemble those in the cells of today's simplest land plants, such as the mosses and liverworts, than do the structures in multicellular green algae. If these one-celled algae proliferated, the multicellularity may have arisen independently on land.

Nobody knows what the first land plants looked like, but their survivors are generally considered to be today's bryophytes: mosses, liverworts, and related plants. Bryophytes have only a primitive system of tube cells for conducting water—not the true vascular plants' xylem cells with lignin-thickened walls. Thus they are not fully emancipated from a moist environment.

The sperm cells of bryophytes have flagella; to reproduce, the sperm must swim through a film of water over the plant's surface to reach the egg cells. Chemical attractants diffused from the egg guide the sperm. Bryophytes do, however, have the cuticle common to land plants, an outer layer of cells specialized to retard evaporation from the plant's interior. The cuticle also shields the plant from ultraviolet radiation and contains chemicals that repel herbivores and microbes.

Although it took plants some three billion years to begin their assault on the land, the evolutionary opportunities in the virgin realm were diverse and plant evolution fairly raced. The next 100 million years saw the evolution of elaborate vascular systems; for complex plants, an early need was the movement of roots and metabolic products down from the photosynthesizing tissues (the earliest land plants lacked true leaves) to the roots and other tissues that could not make their own food.

The tendency toward treelike forms, arborescence, also began in these early plants. Height has obvious survival value where plants compete with one another for access to sunlight, and in virtually every group of land plants treelike forms have evolved. For example, during the coal age, or Carboniferous period, some 300 million years ago, there were great forests of plants resembling hypertrophied versions of today's club mosses and horsetails. In this comparatively brief period, plants evolved true roots, branching stems, and leaves.

In this period there was another major step in evolution: the invention of the seed, which amounted to yet another adaptation for life on dry land. A seed is a plant embryo encapsulated, along with a small food supply, in a coat that protects the contents from desiccation and decay. Thus sealed, the embryonic plant can rest dormant for weeks or years, sprouting only when conditions are favorable.

The seed habit

Seeds with hooks can hitch rides on passing animals and become widely dispersed. Seeds inside a fleshy fruit can be eaten by animals and excreted along with a supply of fertilizer in distant places, the seed coat having resisted digestion in the animal's gut. Windborne seeds may have an even wider range, with deposition in diverse environments that make adaptation and evolution more possible.

Predecessors of the seed plants could complete their life cycle only in a fairly moist environment. Take the ferns as an example. Like all their ancestors, ferns alternate generations. The stage represented by the familiar leafy fronds is diploid, possessing paired chromosomes. Through meiosis, the fern produces haploid spores, each with unpaired chromosomes, on the undersides of its leaves. These spores fall to the ground and grow into tiny haploid plants—the gamete-bearing half of the fern life cycle and the one of which most people are completely unaware. The spores grow into heart-shaped sheets of green cells, plantlets each less than a centimeter across. They put out rootlike hairs, called rhizoids, that tunnel into the ground for water.

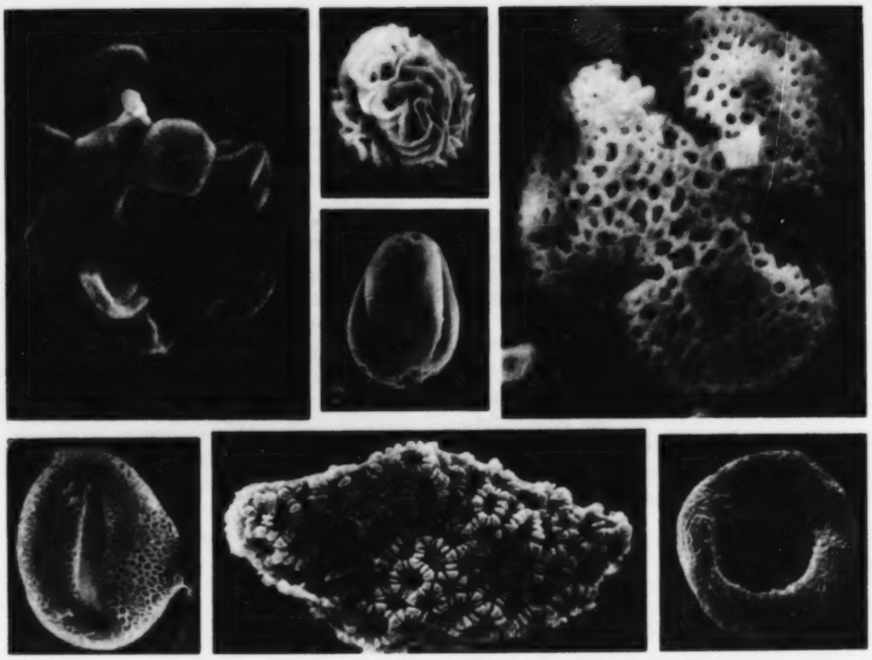
This inconspicuous haploid plant produces both male and female gametes. The sperm are released and, as with the bryophytes, must swim through a film of water across the leaflike parent body to reach the egg cells. The resulting zygote becomes an embryo that must root quickly before it dries out. Then the more familiar form of the fern can grow. Without a moist environment, ferns cannot reproduce.

The transition to what botanists call the seed habit, a stratagem that let plants move into drier environments, is believed to have occurred initially among a now-extinct group of plants, the seed ferns. William H. Gillespie of the United States Geological Survey, Gar W. Rothwell of Ohio State University, and Stephen E. Scheckler of Virginia Polytechnic Institute and State University recently discovered the fossil remains of the earliest known such seeds in late Devonian deposits (dated to about 350 million years ago) in the coal fields of West Virginia.

The seed ferns—a major component of Carboniferous era forests—were the first of

---

Rensberger is a senior editor at Science 82.



Fossil pollen. The oldest recognizable angiosperm pollen grains (bottom three) have single apertures for pollen tube germination, a gymnospermlike feature retained by relatively few living angiosperms (e.g., magnolia); grains with three apertures, as in most modern angiosperms, appeared soon after. By the end of the early Cretaceous, the grains had become varied in size and shape (top four). The total diversity of angiosperm pollen types, however, was still only a fraction of that seen today. (Magnification about 1,300 times)

James A. Doyle

the gymnosperms. This is the more primitive class of seed plants which, in later times, would give rise to the fernlike cycads and the conifers.

All gymnosperms rely on the wind to carry their pollen, the immotile equivalent of the sperm of more primitive plants. The more advanced class of seed plants are the angiosperms, or flowering plants, which attract insects or other animals to ferry pollen from plant to plant. (The names of the two groups of seed plants refer to whether the seed is enclosed in an ovary or fruit: gymnosperm means naked seed and angiosperm means enclosed seed.)

The key step in the origin of the seed habit was to stop shedding spores. Instead they would be retained as a kind of parasite on the plant, which already has deep roots reaching subsurface moisture. This is what all seed-bearing plants do today.

A related innovation was to make two kinds of gamete-bearing plants, male and female. The pine tree is a good living example. Each tree makes two kinds of cones, male and female. At the base of each scale on a female cone are two small plants, each bearing eggs. The male cones have sacs that burst, casting the dustlike pollen to the wind. Of the millions of pollen grains released, only a few chance to be blown into a female cone where they can fertilize an egg.

A better way

Given the apparent wastefulness of wind pollination, Cornell's Niklas has looked for evolutionary trends that might have occurred to improve its efficiency. He has found two possible examples. The first involves the integument around the seed's egg-bearing ovule.

In the early 1960s, H. N. Andrews, now of the University of Connecticut at Storrs, speculated that the variety of seed types found among fossils of the seed ferns represented an evolutionary sequence. In the putative earliest form, the ovule was naked, circled only by a sort of picket fence of tiny green branches. According to Andrews, these pickets eventually fused, enclosing the ovule and leaving only a tiny hole at the top for a pollen grain to get in. This is the form all seed plants have today.

Rather than concealing seeds from pollen, Niklas has found, this process of enclosing the seed would have improved the ability of windborne pollen to reach the ovule inside. He has built scale models of the various primitive ovules and put them in a wind tunnel, where they were exposed to different wind conditions; helium-filled soap bubbles were used to make visible the airflow around the models. While the wind carried the bubbles past the model ovules, as pollen might be carried, Niklas videotaped the action.

In the most primitive ovule model, the one surrounded only by the open picket fence, the bubbles blew past, being deflected slightly by the turbulence created when the wind hit the fence. Relatively few bubbles hit the ovule. When Niklas substituted models of the intermediate ovule form, in which the pickets are partly fused and enclose part of the ovule, the videotape actually shows some bubbles flying past the ovule, getting caught in a backdraft, and being pulled toward it.

In the most advanced ovule the pickets have fused almost completely, leaving only a tiny opening at the top. There the wind currents actually catch bubbles and pull them into the opening; there is a high incidence of bubbles hitting the ovule, especially near the opening, and actually entering it. "This is the first time," Niklas says, "that the aerodynamic attributes of plant shape have been related to plant function."

Niklas concedes that his wind-tunnel experiments do not prove that the modern form of the seed evolved to enhance its pollen-catching ability. But he notes that evolution would not be likely to have followed a path that reduced the efficiency of reproduction.

Recently Niklas has taken his studies of wind pollination further by testing the aerodynamic properties of pine cones. Conifers are by far the most widespread of the gymnosperms. They arose in the Carboniferous and have been an important group ever since. Today's dominance of such gymnosperms as pines and firs in the world's northerly latitudes suggests that their method of pollination must have something going for it.

One of those things, Niklas has found, is the shape of the pine cone itself. Again putting scale models in wind tunnels, he has shown that each scale acts as a tiny airfoil, sucking pollen grains into the cone. The female pine cone is a kind of black hole for pollen.

The flower habit

Efficient as wind pollination may be, it has limitations. For one thing, it is effective mainly where the plants grow thickly over a large area. Widely scattered gymnosperms would have a poor chance of being pollinated by a faraway mate. Even in a pine forest, however, the trees must produce huge quantities of pollen to ensure adequate fertilization. When the cones are shedding pollen, the floor of a pine forest can be blanketed several millimeters deep with lost pollen. Producing such quantities represents a high energy cost for gymnosperms and leaves the way open for a more efficient system to evolve.

And one did, some 300 million years after plants reached land, between 130 million and 110 million years ago. It was during the Cretaceous period, when dinosaurs still roamed the land, that a way of improving on the gymnosperms evolved. The world received its first flower.

The invention of the flower, which involved several modifications in the structure and behavior of the reproductive organs contained within the blossom, was achieved in concert with changes in the vegetative, or nonreproductive, tissues. Of these, one of the most important was the acquisition of an extraordinary architectural plasticity that has allowed angiosperms to grow into a bewildering variety of shapes.

Within a few million years, these new features triggered a major adaptive radiation that propelled the angiosperms into dominance over almost all the earth's land. From those first flowering plants we have roses and rice, oaks and onions, palms and pansies, magnolias and mangos, and most of the other familiar plants of today's world. Not all have the showy petals commonly associated with the word flower; the inconspicuous pollen-bearing structures of many trees and the tassels and ears of corn are also, technically, flowers.

The flower, essentially, is a plant's way of enticing small animals, chiefly insects, to collect a little pollen from one plant and carry it to another. Flowering plants can be widely spaced because the insects are highly mobile and, in many cases, seek out only flowers of a single species.

Abominable mystery

In many cases, there is evidence that the evolution of the pollen producer and the pollen distributor was closely related. (See "'Feedback' Produces a Theory of Ecology," *Mosaic*, Volume 10, Number 6.) Coloration and scent of specific flowers attract specific insects. In some highly specialized cases, more than scent and color are involved; in orchids, for instance, the flower's reproductive parts resemble females of the pollinating insect, and males actually try to inseminate the blossom.

The origin of the flowering plants has long been one of the world's great evolutionary puzzles. Darwin called it "an abominable mystery." Recently, however, the mystery has moved closer to solution through the work of Ledyard Stebbins, Leo Hickey, James A. Doyle of the University of California at Davis, and A. L. Takhtajan in Liverpool. These scientists are far from full agreement with one another, but among them, they agree, progress is being made.

Hickey says one of the problems that frustrated earlier workers was the tendency of 19th-century botanists, when identifying fossil leaves of early angiosperms, to give them the names of the modern groups they most closely resembled. This created the impression that when the angiosperms first appeared in the fossil record, they had already evolved into many of the modern groups. Thus Darwin had been led to speculate that the angiosperms must have arisen and diversified on some now-lost continent from which they all spread to the known regions. Hickey, who specializes in fossil-leaf architecture, says it is now clear that the leaf resemblances are only superficial.

Stebbins has interpreted the fossil evidence of early angiosperms to suggest that they began as small, rapidly maturing shrubs adapted to upland habitats with seasonal dry spells. These would have been habitats not already dominated by the slow-growing gymnosperms.

Hickey and Doyle speculate that because of their short life cycle, early angiosperms would have found it easy to spread into any area where natural processes frequently disturbed the soil. That idea grew out of their analysis of early angiosperm fossils found preserved in sediments laid down along the margins of ancient streams.

In effect, they say, the ancestors of today's flowering plants got their start as weeds. These two, who work closely together but don't always fully endorse each other's conclusions, agree that their new interpretation of angiosperm origins tends to rule out the old textbook view that the earliest flowering plants were treelike forms such as the magnolia. It now seems more likely, they say, that the angiosperms—like so many of their predecessors—independently invented the tree.

Takhtajan has suggested that the means by which slow-growing gymnosperms evolved into fast-growing angiosperms was neoteny, a process invoked by one school of evolutionists to explain many evolutionary changes. Neoteny is the condition in which an individual's normal period of development is cut short at a juvenile stage; neotenic species reach sexual maturity when they still look like the juvenile forms of their ancestors. Takhtajan and some others are led down this road by the similarity between some of the reproductive organs of angiosperms and immature forms of gymnosperm organs.

Counterrevolution

Once the flowering habit was established, its propensity for rapid speciation and adaptation eventually produced longer-lived, more treelike forms. These could expand beyond transitory habitats and compete successfully with gymnosperms. Thus many of the later flowering plants came to be as slow growing as the gymnosperms. "In a sense," Karl Niklas wrote in a paper with Bruce H. Tiffney of Yale University and Andrew H. Knoll of Oberlin College, "the angiosperms behave as good revolutionaries, initially toppling the old regime, but ultimately usurping its manners and life-style." In fact, some of the angiosperms have even reverted to wind pollination. The grasses—arguably the most important group of plants from a human perspective because they supply all the food grains—are a prime example.

The angiosperms arose only within the last third of all the time plants have been on land and the most recent 4 percent of all the time life has existed. But they have adapted to so many environments and ways of life that they now number some 275,000 known species, five times as many as all other kinds of plants put together.

From a zoocentric point of view, plant evolution may not seem to have been quite as astonishing as animal evolution. For one thing, plants lack the complex behavioral systems that characterize animals. And, in a sense, one may regard all the green plants simply as variations on a theme—they are all just ways of capturing sunlight.

And yet, in a larger sense, it is clear that animals and plants have participated in a coevolutionary process. Where there are no plants, few animals can live. More than anything else, it is the various kinds of plant communities that, by making available different kinds of food and cover, define animal habits and habitats. Without plant evolution there could have been no animal evolution. And there is surely much more to learn about the development of life on earth—animal as well as vegetable—through a more intense study of the evolution of plants, which made it possible. •

The National Science Foundation supports part of the research discussed in this article through its Systematic Biology Program.

Integrated Optics: A Communications Revolution

by Mort La Brecque

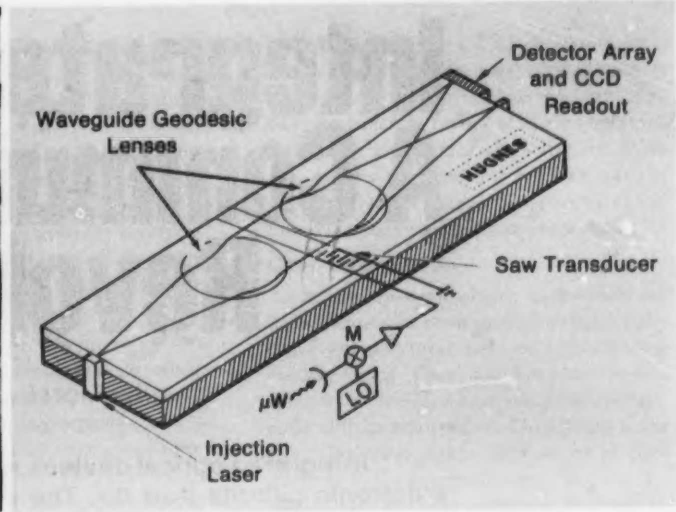
Integrated optical devices will do much that integrated electronic circuits now do. The problem of the optical switching transistor may set the limit.

For a variety of reasons, integrated optical circuits did not match the high expectations held for them in the 1970s. They are likely to in the mid-1980s, however, and there appears to be a neck-and-neck race for first practical application. That device will probably be either a spectrum analyzer for the military or a device for single-mode fiber optics communication. In any case, the technology will have its greatest impact on the world of telecommunications.

Antiaircraft missiles are designed to home on target aircraft with radar. An effective response can be made by the target: jamming the enemy signal so that the tracking is stopped or launching one's own missile to destroy the enemy's. First, though, the signal must be identified—no easy task if the plane is flying over a densely populated region; enemy radar is then buried in an electronic amalgam that includes television, microwave communications, and other radiation.

There are radio frequency spectrum analyzers that do this seemingly impossible

Bell Labs



job. Every two microseconds the analyzer sorts signals into 100 frequency bins, each covering a small part of a 400-megahertz electromagnetic spectrum, while also measuring the signals' amplitudes. The data are passed on to a computer aboard the plane, which reports its analysis to the pilot. From the signals' amplitude and frequency, it can be determined whether a source is homing, and if so, at what speed.

The electronic spectrum analyzers aboard modern military aircraft are effective but take a whole rack of digital equipment—considerable weight and valuable space on an airplane. A dramatically miniaturized version, no larger than a stick of chewing gum, may make history when it debuts—possibly in the 1980s. It will not be electrical, but optical, and it is neck-and-neck with a civilian communications device in a race to become the first integrated optical devices ever to make it from the laboratory to the world of technology.

Integrated optics is the optical analog of intergrated electronics—miniaturized versions of bulk devices mounted on a common substrate. There are important differences, however, which suggest that the newer technology will complement, not supplant, the older one.

For instance, although integrated optical circuits are only a few millimeters long and a few microns wide, miniaturized optical devices are unlikely ever to approach the level of density possible with electronics. Very large scale integrated electronic devices today contain hundreds of thousands of transistors and may in the near future contain millions. (See "Faster Switches, Smaller Wires, Larger Chips," *Mosaic*, Volume 13, Number 1.) Integrated optical devices in development, although smaller than their electronic counterparts, typically have no more than ten components. Since the laws of physics mandate that one key component—the semiconducting laser—must be of a minimum size to function properly, integrated optical lasers can get somewhat smaller than they are now (between the size of a matchhead and that of a grain of sugar) but will never achieve the miniaturization of a transistor in an electronic circuit. (See "Optical switching transistors" accompanying this article.)

Hybrid solutions

Nor are most integrated optical devices really integrated in the sense that electronic circuits are: monolithic, self-contained units made of one material. Such monoliths are highly desirable because of the increased reliability, reduced signal loss, and greater circuit complexity they offer.

Semiconductors are the materials of choice for integrated optical monoliths, since they are the only materials known that can change an electrical signal into an optical one, and vice versa. Electronics uses them—one-element silicon or two-element gallium arsenide. In addition to a two-element compound, optical semiconductors contain three- or four-element compounds, making them ternaries or quaternaries in the language of the trade.

Each component of an integrated optical monolith must have its three or four elements in a unique proportion in order for it to perform its function. This increases dramatically the number of steps in the fabrication process. Until very recently, fabrication technology has not been able to provide the control necessary to go through each step with sufficient precision, and the chances of producing a perfect optical monolith were infinitesimal.

New modes of fabrication improve the prospect of making integrated optical circuits on a practical scale. But as Amnon Yariv of Caltech points out, most miniaturized optical devices are not integrated. Dielectric materials, insulators rather than conductors, do a better job than semiconductors in guiding lightwaves from one optical component to another. To reflect the importance to the research field of thin dielectric waveguides, Yariv would like the field renamed integrated and guided-wave optics.

Hybrid devices have almost uniformly been the pragmatic answer to the design problem, using each kind of material for what it is best suited to do. The spectrum analyzer, for example, is a hybrid device with a waveguide and active optics made of a dielectric material coupled at one end to a semiconductor laser and at the other to a semiconductor detector.

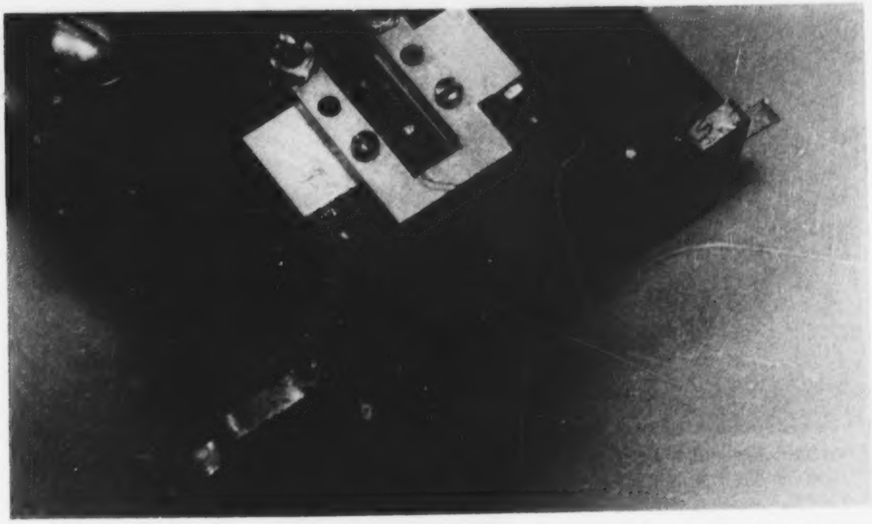
In some respects, integrated optics are superior to electronics. They are, for example, immune to electromagnetic interference from lightning, motors, and other sources of electricity. If two optical waveguides are placed close together, there will be very little crosstalk—signals leaking from one guide to the other. The signal-carrying capacity of an electromagnetic wave increases with its frequency. Since lightwaves are typically 20,000 to 200,000 times as high in frequency as the microwaves that traverse electronic devices, data can be transmitted much faster optically.

Parallel processing

But it is in signal processing that devices like the integrated optical spectrum analyzer really win the race over their electronic counterparts. Optical spectrum analyzers work like another optical device, the camera, processing all the frequency components of a

---

La Brecque is former editor of The Sciences.



Spectrum analyzers. Seen here are various versions of the radio frequency spectrum analyzer: Hughes Research Laboratories' version and a schematic (opposite page); competitive Westinghouse Advanced Technology Laboratory version of the spectrum analyzer (above), showing a closeup of the laser (tiny rectangle with a wire touching it); Chen Tsai's spectrum analyzer (below) uses an acousto-optic Bragg cell device.

Hughes Research Laboratories, Westinghouse Advanced Technology Laboratory, Chen Tsai/University of California at Irvine

signal in parallel, simultaneously. Electronic spectrum analyzers are sequential processors. They do their job point by point, just as an electron beam scribes a picture across a television screen. This accounts for the dramatic difference in size of the two devices.

The key building block of integrated optics is the thin-film waveguide, made from a transparent material a single micron thick. Total internal reflection is the basis of waveguiding; a light ray will zigzag back and forth through the thin-film waveguide, being kept on course by undergoing successive total reflections at both its upper and lower interfaces with the surrounding media: usually the air above the waveguide and the substrate on which the waveguide is fabricated. All that is required is the processing of the waveguide so that it has the greatest refractive index—the velocity of light in a vacuum divided by its velocity in the material.

The waveguide may be planar, extending across the entire surface of the film. A more versatile form is the channel waveguide, which confines the lightwave on either side. Channel waveguides can be made as straight and narrow strips or as curves to change the direction of propagation of the lightwave.

Except in the case of monolithic devices that would normally incorporate a laser as an integral part of the chip, light must be coupled into the waveguide. If the light source is a laser adjacent to the waveguide, coupling can be accomplished through a prism without any serious loss of signal. However, prism coupling is a laboratory technique unsuitable for a functional device and cannot be used in any case if the light is emanating from an optical fiber.

Aligning the center of a circular fiber with an elliptical waveguide atop a microscopic chip is a tricky business. As recently as two years ago, 90 percent of the light was being lost in such coupling, casting a dark shadow on the future of waveguiding. But this coupling problem may have been solved with new V-channel couplers, developed independently by a number of groups. Bell Laboratories' Stewart Miller calls Bell's 80 percent transmission (from a single-mode fiber through a centimeter-long integrated optics waveguide and back into a fiber) the "lowest published loss value," but Adrian Popa of Hughes Research Laboratories says that Hughes's couplers transmit 90 percent of the light from fiber to waveguide.

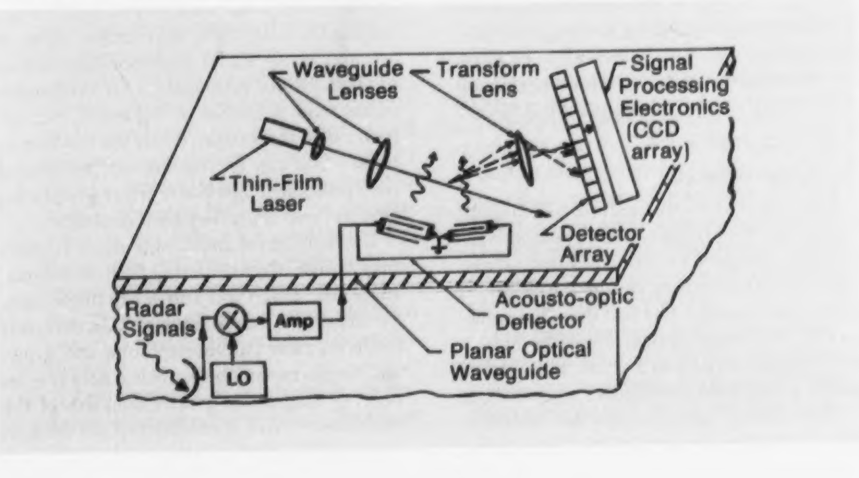
Electro- or acousto-

Couplers, lenses, beam splitters, and such are passive parts of integrated optical systems. There must also be active components. These manipulate the lightbeam as it travels along the waveguide by temporarily changing the guide's refractive index. This can be done in two ways, depending on the waveguide's properties: with electrically generated sound waves (acousto-optically) or with an electrical field directly (electro-optically).

A signal can be switched electro-optically from one waveguide to another with the directional coupler, a device in which identical waveguides approach each other, run parallel for a short distance (the coupling region) and then separate. Ronald Schmidt of Xerox's Palo Alto Research Center explains such a coupled-mode interaction this way: "In physics classes, they present the example of two pendulums connected by a spring. If you make one vibrate while the other is at rest, the second will begin to vibrate and the first goes to rest. Analogously, as light in one waveguide enters the coupling region, it gradually transfers to the second waveguide, staying there as the guides separate."

To keep the signal in the original track, electrodes are placed over the waveguides in the coupling region. If applied voltage is directed oppositely in each guide, the linear electro-optic effect causes the refractive index of one to increase and of the other to decrease. Light propagating in the first guide no longer has an identical guide next to it and doesn't couple over.

An acousto-optic equivalent of the directional coupler (a Bragg cell) has been devised by Chen Tsai of the University of California at Irvine. The device is a media go-between that, in response to high-frequency electronic signals, generates broad, fan-shaped beams of surface acoustic waves. These are inaudi-



ble waves of the same high frequency whose energy is concentrated in a thin layer just beneath the surface. The waves propagate through a planar waveguide made of an acousto-optic material. By continually rippling the waveguide's index of refraction—the ripples depending upon the acoustic frequencies present at any instant—the sound waves in effect form moving optical gratings. [See "Slower Can Be Better," *Mosaic*, Volume 8, Number 3.]

Guided light waves, traveling in a direction such that they intersect the sound waves at a particular angle, encounter the gratings and are diffracted by them. The gratings simultaneously deflect different small portions of the light—less than 1 percent each—at slightly different angles from the main beam. The angle of each deflected beam is a measure of a frequency component of the original signal; the intensity of each beam is a measure of the frequency component's amplitude.

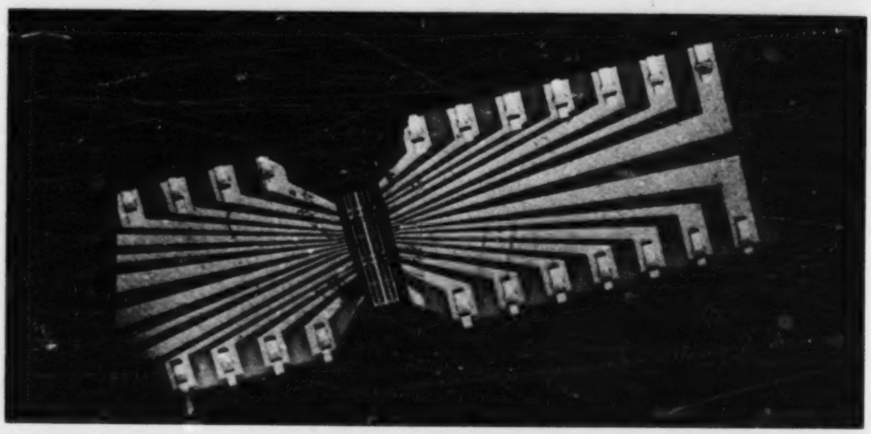
The radio frequency spectrum analyzer has a Bragg cell at its heart. Other components include a laser to generate the lightbeam, a lens to adjust the beam before it encounters the optical grating, another lens to focus the minibeams diffracted by the optical grating, and a photodetector array to receive the diffracted lightbeam and produce equally intense electronic signals in the same range.

The concept of an integrated optics spectrum analyzer was first suggested in the early 1970s by Michael Hamilton and his colleagues at the Air Force Avionics Laboratory. Hughes Research Laboratories was awarded the development contract by the Air Force, and shortly thereafter the Naval Research Laboratory awarded another contract to Westinghouse Advanced Technology Division. The companies have each built their own versions of the analyzer.

Both work, but so far both fall short on one critical specification: dynamic range, the variation in amplitude to which the device will respond. "We have to be able to see at the same time two signals that differ by maybe even a million times in amplitude," says Westinghouse's Dickson Mergerian. Otherwise, he explains, a very strong radar signal could be transmitted to hide a very weak one.

The main line

Perhaps the greatest use of integrated optics will be made in telecommunications. Phone companies around the world are beginning to switch from copper coaxial cables to optical fibers in their transmission networks. The Bell System has 10,000 kilometers of fiber in service today, mainly for intracity communications between central offices, but will link cities optically—within the Northeast and along the West Coast—by the mid-1980s; further in the future are plans for a transatlantic fiber cable.



Prototype. The ribbed lightguide structures of this prototype integrated optical circuit are used to switch lightwaves.

Bell Labs

The optical components being used with this growing fiber network are all discrete, just as the first transistors were. But an active program aimed at combining discrete components in optical chips is well under way. According to Joseph Giordmaine, director of Bell Labs' electronic and photonic technology laboratory, "It's already apparent that telecommunications needs a number of integrated optics components and devices."

It is appropriate that Bell Labs should have an active integrated optics program, since what is generally regarded as the first vigorous pursuit of the concept began with Bell's Stewart Miller in 1969. [The idea itself is usually credited to R. Shubert and J. H. Harris, who published it in 1968.]

The antecedent of integrated optics was the laser, according to Miller, now director of Bell's guided wave research laboratory. "We wanted to use a multiplicity of lasers and transmit their beams between cities as a very large volume communications medium," said Miller. "In order to make that feasible, we needed radically different techniques from what was then the art for handling optical waves."

On again, off again

Electronic miniaturization had begun, suggesting that lithographics could also reproduce complex miniaturized optical circuits in large numbers. But how well would these circuits work? The first thin-film waveguides had huge optical losses at only a millimeter or so of length. P. K. Tien, one of Miller's Bell Labs colleagues, was the first to demonstrate optical chips that wouldn't lose most of the power injected into them.

Ironically, a similar development in optical fibers at about the same time slowed down progress in integrated optics. Although optical fibers had been used in specialized transmission applications since the 1950s, they were not considered serious contenders for long-distance communications because the power of the signal propagating in them was halved each meter it traveled. When it was discovered that the high losses were caused by impurities introduced in the fabrication process, the impurities were removed and the first low-loss fibers produced.

"Integrated optics then no longer appeared so necessary for optical communications," recalls Miller, "since semiconductor lasers could be coupled directly into the fiber without complex circuitry. Before the low-loss fibers came along, we planned to transmit light along guided paths with lasers and mirrors. In that discarded concept, we had to have integrated optics."

Further developments in optic fibers, however, have helped improve the prospects for integrated optics in telecommunications. Fibers currently in use are multimode: They propagate only light that enters them at certain discrete angles corresponding to their modes. Pulses propagating in multimode fibers tend to spread in duration. This effect, called dispersion, limits the number of pulses that can be transmitted per second (two pulses sent too close together would soon overlap) and limits the fibers' data rate.

Recently developed single-mode fibers—they transmit light entering them at only one angle—are much less subject to dispersion. Eventually, the data-transmission demands made on fiber communications will grow, and single-mode fibers, with a data rate an order of magnitude greater than that of the multimodes, will gradually replace them in most applications. Integrated optics is a

Optical switching transistors

Will complex switching operations remain entirely in the domain of electronics for the foreseeable future? Some observers are more sanguine than others about the promise of all-optical switches. Formidable obstacles exist to making such devices. But the recent discovery of certain optical materials that can respond at least one hundred times as fast to a signal as can the fastest electronics has encouraged some physicists and electrical engineers.

They look forward in particular to the day when the logical switching operations that are the heart of digital computation will be done by an optical version of the switching transistor. Using light to switch light, if practical, would make possible computational devices of unimaginable speed. For this to happen, however, nonlinear optical systems must first be developed.

Conventional optical systems are linear: Light injected into them will slow down at a constant rate over distance. In a nonlinear system, injected light would slow down at a rate dependent on the amount of light present; the more light injected into the system, the more it would be slowed down. The input-output relationship would not be linear.

One of the basic logic circuits of digital computers, the exclusive or gate, requires nonlinear response. Switching transistors can perform this kind of operation because they have two different stable states, a condition called bistability.

A bistable optical device, an optical equivalent of the transistor, was first proposed in 1970 and first demonstrated switching in 1975. It has excited enthusiasm in some researchers and skepticism in others. Its two stable output states are two different light intensities; which one switches on depends on the "history" of the input light's intensity.

As Ronald Schmidt of Xerox explains: "If you put low-intensity light into the device and gradually increase it, you'll continue to get the same intensity out. At a given intensity input, however, the output intensity will increase dramatically. If you reduce the input intensity, the output will stay in a high-intensity regime until another particular input intensity is reached; the output will then switch down to the original intensity level."

Two versions of the bistable optical device are being investigated. One is an all-optical mechanism that produces its effect intrinsically, the input light inducing changes in the material out of which the device is made. The second, a hybrid version, uses an external component to make such changes; part of the input light is sent to a photodetector, which converts it to an electrical signal that is fed back to the device.

The speed of the hybrid device is limited by the speed of its electronic component, so it offers no real advantage over all-electronic logic switches that might be built in gallium arsenide. (Integrated circuits using gallium arsenide operate at ten times the speed of their silicon counterparts.) "The hybrid device can be built right now," says Michael Barnowski of TRW Technology Research Center, "so it gives us a way to study the phenomenon of optical bistability. But the all-optical version is that one that could produce exceedingly high data rates."

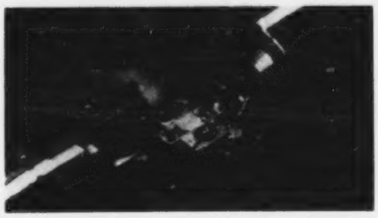
The first discrete (nonintegrated) optics version of a bistable device, which used sodium vapor as the nonlinear material, was bulky and switched states very slowly. A much more practical one has now been demonstrated. Hyatt Gibbs of the University of Arizona's optical sciences center reports that by going to gallium arsenide for nonlinearity, he has reduced the size of the device to microns, the time for the switching-up operation to picoseconds, and the time for switching down to nanoseconds—all at room temperature.

An all-optical bistable device has been worked out in theory by Elsa Garmire of the Center for Laser Studies at the University of Southern California. Peter Smith of Bell Labs has made a working integrated optics version containing a hybrid electrical feedback device. It has demonstrated only a single operation, however. A large array of logic gates made of many such devices would be required to do something as simple as addition.

Optical bistability is still in the early stages of research. More work is being done on theory than on building experimental devices, and it is unknown when, if ever, a technological payoff will occur. There is general agreement, however, that initially it would not be in the form of a general-purpose computer.

The basic problem with the semiconductor materials now being considered is that their nonlinear effects require large amounts of optical power for them to be driven at optimal speeds. "When you start dissipating power," says Henry Taylor of the Navàl Research Laboratory, "you have to remove that power. If you're using them by the thousands in logic elements in a very high speed computer, the power dissipation per unit area grows increasingly large, quickly reaching the limit possible with any known heat transfer agent."

An all-optical computer that used less power would switch them more slowly than do the integrated circuits of electronic computers. Speed considerations also militate against less densely packed optical logic gates driven with high power: The time lost in transit between them would be greater than the time saved in switching optically.



Optical transistor. Bell Labs is investigating this hybrid bistable optical device (optical switching transistor).

Bell Labs

A more likely first use for optical logic devices is on a smaller scale—as switches for optical communications. "The carrier is already in optical form," says Taylor, "and if you can process the signal without converting to electronics, the system design is that much simpler."

The heat generated by a small number of bistable optical devices, switching at much slower speeds than would be required in a general-purpose computer, is not excessive. However, optical powers used in telecommunications are well below the level needed by bistable optical devices to perform useful logic functions.

Elsa Garmire points out that for different applications, such as image processing, light may do a better switching job than electronics. "We need new nonlinear materials," she says, "that require less power to switch states at speeds beyond nanoseconds." Nothing demonstrated so far looks like it can do that, Garmire adds, but many avenues remain unexplored. Perhaps one will lead someday to the right stuff for bistable optical devices. •

single-mode technology that can increase the new fibers' transmission capabilities even more.

Expanding pains

Despite the promise of integrated optics, and more than a decade of research and development, no such devices are yet in use in telecommunications or any other field. This apparently lethargic growth rate is explained by Ronald Schmidt, who points out the difficulties involved in creating a reliable working device.

"Technology's much like a balloon," says Schmidt. "You take a device and push it out in one direction, and that says high speed. Then you do another experiment with a different device and push it in the direction that says low voltage. Then you push out a third device in the direction of high coupling efficiency. You've demonstrated that you can do all those things in different devices, but before any one of them can be put into a system, you have to be able to push it in all directions at once and expand the whole balloon."

The key to such an achievement, according to Bell's Stewart Miller, is materials studies. Glass, the first materials system investigated for waveguides, has very low optical losses. Unfortunately, the electro-optic effect is absent in glass, limiting to passive components its use in complex integrated optic circuitry.

Unlike glass, lithium niobate has both excellent electro-optic and acousto-optic properties, and it can be used for active as well as passive components. The spectrum analyzer's acousto-optical circuitry, for example, is made of lithium niobate, and the material's electro-optical efficiency makes it the usual choice for directional couplers.

But lithium niobate has its limitations: Its dielectric nature prevents it from conducting electricity very well and thus from being used to make a laser or detector. For these two fundamental components of integrated optics, which also provide interfaces with electronic circuitry, one must turn to semiconducting materials.

Honing the laser

Semiconductor lasers have many attractions, being of a size compatible with waveguides, operating from low voltages, and using no moving or fragile parts. Invented in 1962, they were first made entirely of the single compound gallium arsenide.

That uniform homostructure could not stimulate lasing action unless it was pumped with very high currents at very low temperature. Improved performance was obtained by dividing the laser into two layers—making it a heterostructure—of gallium arsenide and gallium-aluminum arsenide. But it wasn't until the double-heterostructure laser was invented that semiconductor lasers became practical. Made of a layer of gallium arsenide sandwiched between two layers of gallium-aluminum arsenide, the first double-heterostructure laser (unlike earlier semiconductor lasers) could be operated continuously at room temperature.

Used with silicon detectors to convert their optical signals to electric ones, gallium-aluminum arsenide ternaries—three-element semiconductors—are excellent, long-lived lasers. But they have a fatal flaw for long-distance fiber communications. They transmit signals at short infrared wavelenths—about 0.85 micron. The single-mode fibers and waveguides that will dominate optical communications, however, operate at longer wavelengths, between 1.3 and 1.5 microns, where attenuation is lower.

When it became clear that the 1.3- to 1.5-micron range was optimal for long-distance fiber waveguiding, researchers began to look for materials that would produce semiconducting lasers operating in that range. A number of ternary and quaternary compounds were tried in various double-heterostructure arrangements. The best proved to be a quaternary, first demonstrated at MIT's Lincoln Laboratory in 1976, in which a layer of gallium-indium-arsenide phosphide was sandwiched between two layers of indium phosphide. Its wide operating range, from 1.0 to 1.6 microns, permits transmission of many more channels via wavelength division multiplexing—sending different channels of information on adjacent wavelengths—than earlier semiconductor lasers.

All that was then needed was a detector to convert the optical signals put out by that laser to electrical signals. Such a semiconductor detector was developed in 1978, out of indium-gallium arsenide, insuring the hegemony of the four-element laser in optical communications in the coming years.

Producing monoliths

Semiconductors can be made into electronic devices as well as lasers and detectors. The fact that the same material can be used in both electronic and optical components suggests to Caltech's Amnon Yariv the possibility of combining the two in what he terms opto-electronic devices, which could be the interfaces between electronics and optics.

Yariv has built an opto-electronic repeater to amplify optical signals weakened by their long-distance travels over optic fibers. The world's first such device, it is a monolithic structure whose major components are three transistors and a laser made of gallium arsenide.

Semiconductors must be used in complex monoliths, since they are the only materials from which all fundamental-device components can be made. Up to now, semiconductor lasers have been made by a process called liquid-phase epitaxy. (Epitaxy is the oriented layering of a material on a substrate that dictates the orientation.) This process, according to Joseph Giordmaine, "can't be controlled nearly well enough to build reproducible monoliths that might contain many components in addition to the laser."

Liquid-phase epitaxy is unable to provide side-to-side variations in composition, a necessity for building up an array of devices with different characteristics on a single chip. Metal-organic chemical vapor deposition and molecular-beam epitaxy, two newly developed fabrication processes that grow oriented semiconductor layers out of elements in their gaseous form, can provide the requisite variations. Chemical-vapor deposition is cheaper but molecular-beam epitaxy, because it occurs in ultra high vacuum rather than at atmospheric pressure, offers the ultimate in uniformity.

With the means at hand to build practical working monoliths, researchers have now begun to look at semiconductors for optical components other than lasers and detectors. Fred Leonberger of the Lincoln Laboratory has made directional coupler switches in gallium arsenide, with an eye toward using them as separate modulators for gallium arsenide lasers; such modulators would enable the lasers to operate at very high frequencies, increasing their information-carrying capacity. "If you build this switch on the same chip that the laser is on," says Leonberger, "there is much less loss of optical power between the two devices than if they were separate components."

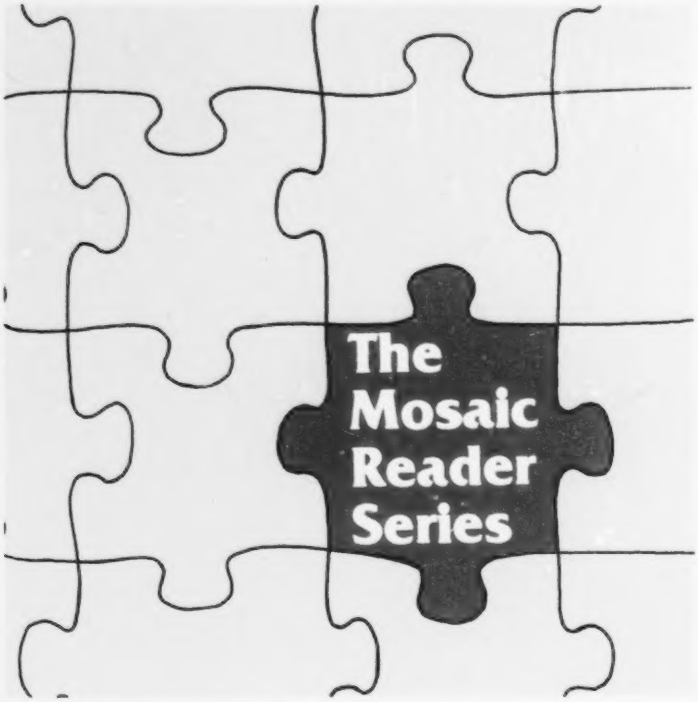
At Bell Labs, Franz Reinhart has actually produced complete monoliths. "Our most complicated device in terms of function," said Reinhart, "is a laser joined together with a ribbed passive waveguide and a modulator. It really touches all the bases."

The technological balloon has not yet expanded uniformly in all directions for monolithic devices, but it is a goal worth pursuing. For only when monoliths can be made on a production line basis will integrated optics at last really live up to its name. •

The National Science Foundation contributes to the support of research discussed in this article principally through its Electrical and Optical Communications, Quantum Electronics Waves and Beams, and Industry-University Interaction Programs.

The National Science Foundation
&
Avery Publishing Group

are proud to present



The genius of the biological scientist

As scientists probe the mysteries within us and our world, **THE MOSAIC READER SERIES** focuses on and highlights current significant research at leading university, industry and government laboratories. This timely and authoritative series will present scientists, educators and students with a concise, intelligent overview of the advances being made by preeminent scientists around the world.

AN AUTHORITATIVE SERIES

The first five books available in this fascinating series focus on research in the specialized fields of the biological sciences. The volumes in this set are:

THE CELL: Inter- & Intra-Relationships
DNA: The Master Molecule
HUMAN EVOLUTION
EVOLUTION: New Perspectives
ECOLOGY: Impacts and Implications

When completed, **THE MOSAIC READER SERIES** will contain more than twenty-four volumes encompassing research in the five major scientific areas: Biology, Anthropology, Physics and Astronomy, Chemistry and Earth Science. This comprehensive reference library of current scientific research will examine topics ranging from the topology of DNA to stellar ontogeny.

Each volume is a collection of articles written and edited by professional science journalists and authenticated by the scientists actually engaged in the research being discussed. These articles, originally appearing individually in *Mosaic* magazine, have been carefully evaluated and specially selected for the breadth and lasting significance of the material they cover. Many of the articles contain insets which highlight topics of special interest that are related to the principal subject. The articles are further enhanced by numerous diagrams, illustrations and photographs. Each volume is indexed and contains a complete glossary.

To order copies of the titles currently available in **THE MOSAIC READER SERIES**, complete the attached order card at the back of this *Mosaic*.

National Science Foundation
Washington, D.C. 20550

Official Business
PENALTY FOR PRIVATE USE, $300

Postage and Fees Paid
National Science Foundation

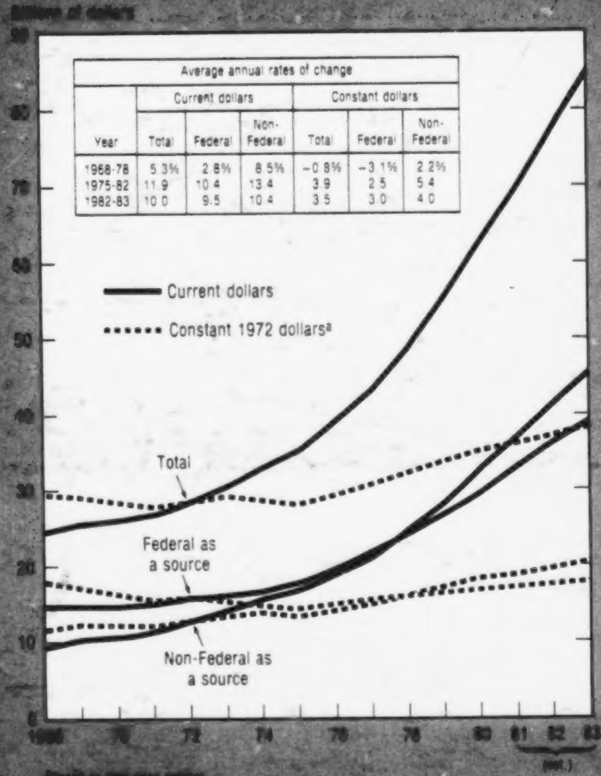
U.S.MAIL

BULK RATE
U.S. POSTAGE PAID
PERMIT NO. G69
WASHINGTON, D.C.

National R/D Expenditures Expected to Reach $85 Billion in 1983

- Total expenditures for research and development in the United States are expected to be an estimated $85 billion in 1983, 10 percent more than was estimated for 1982. In constant dollars,[1] the expected increase between the years is 4 percent. National R/D spending has been increasing in real terms each year since 1975, averaging 4 percent annually through 1983. The growth over this period has resulted from federal increases in defense, space, and energy areas as well as from general industrial increases. It followed a 10-year period of no real-term growth in the nation's R/D effort.
- The federal government is expected to increase its support of research and development activities by nearly 10 percent between 1982 and 1983, to $40 billion. In real terms the increase is estimated at 3 percent. Federal spending of this kind has increased in constant dollars each year since 1975 and is expected to average about 3 percent annually through 1983. In the first five years of this period, government R/D investment reflected the growing national emphasis on energy conservation and development. After that time the emphasis began to shift toward defense activities. The 1983 budget proposals suggest that defense R/D programs will account for 55 percent of the federal R/D total that year, compared to 44 percent in 1980.
- Nonfederal research and development support is expected to increase slightly faster than federal in 1983—10 percent more than in 1982, or 4 percent in real terms—to more than $45 billion. This projected real-term growth is slightly higher than that shown for the previous two years, reflecting hopes that the current recession will end and the inflation rate will go down. Since 1975 nonfederal R/D support has increased in constant dollars at nearly twice the annual rate of federal R/D spending.
- Almost 95 percent of nonfederal support comes from industry each year. The constant-dollar growth in industry spending since 1975 has reflected to a large extent increases in energy and environmental programs.
- National R/D spending as a percentage of the gross national product has increased slightly each year between 1975 and 1982—from 2.2 percent to 2.4 percent; this reflects a real-term leveling of the nation's GNP during this period, compared to a peak of 3.0 percent in 1964. On the assumption that the current recession will end by early 1983, the ratio is expected to remain at the current 2.4 percent level in that year.
- Basic-research expenditures in 1983 should approach $10 billion—7 percent more than in the previous year. National applied-research spending is expected to increase about 9 percent to more than $18 billion. Both basic and applied research have remained relatively level in real terms since 1981. This reflects nonfederal growth as well as federal cutbacks in both areas as industry continues to emphasize longer-range research and the government shifts more toward development activities, primarily in the defense area. National development spending may reach $57 billion in 1983, 11 percent more than in 1982.

National R&D expenditures



Average annual rates of change						
	Current dollars			Constant dollars		
Year	Total	Federal	Non-Federal	Total	Federal	Non-Federal
1968-75	5.3%	2.8%	8.5%	-0.8%	-3.1%	2.2%
1975-82	11.9	10.4	13.4	3.9	2.5	5.4
1982-83	10.0	9.5	10.4	3.5	3.0	4.0

[a] Based on GNP price deflator.
SOURCE: National Science Foundation

---

[1] In the absence of a reliable R/D cost index, the implicit price deflator for the GNP has been used to convert R/D expenditures to constant dollars. The GNP deflator includes the effects of price changes for all goods and services in the economy; it therefore can only indicate approximate changes in costs of inputs specifically related to R/D performance. The increase in the GNP deflator between 1982 and 1983 is estimated at 6 percent.

13/5

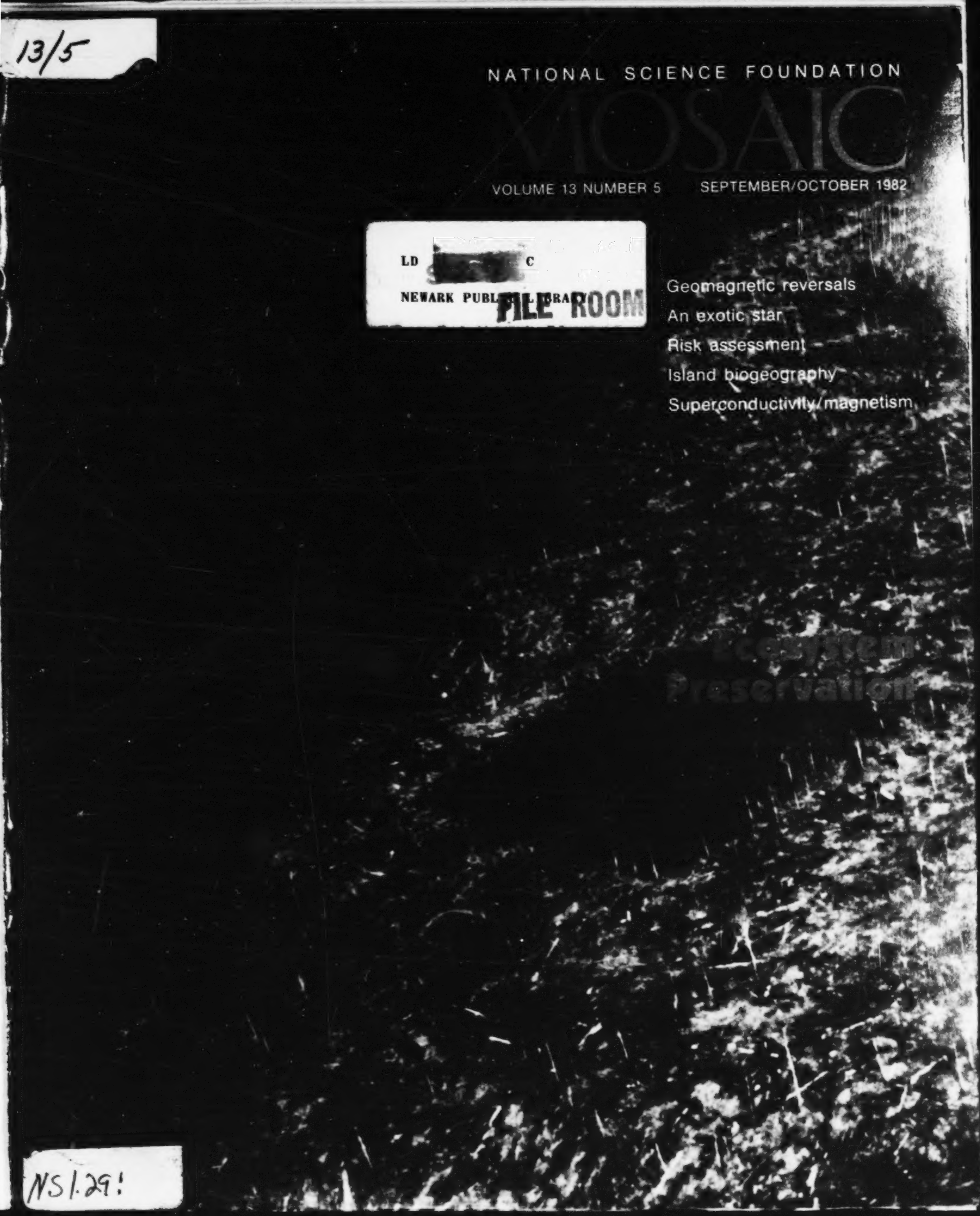
NATIONAL SCIENCE FOUNDATION

MOSAIC

VOLUME 13 NUMBER 5 SEPTEMBER/OCTOBER 1982

LD [redacted] C
NEWARK PUBL[illegible]BRARY
FILE ROOM

Geomagnetic reversals
An exotic star
Risk assessment
Island biogeography
Superconductivity/magnetism



Ecosystem Preservation

NS1.29:

In this *Mosaic*

It seems incredible that a phenomenon as well known as the earth's magnetism is still as mysterious as it is. Most of us, it appears, have accepted for many years a relatively simple explanation: the consequence of rotation of the earth's iron-rich core. This begins to elaborate itself, however, as one looks closer. The core seems to be involved, of course, but there are at least two problems:

- The solid crust of the earth, in which geophysicists have been able to find records of present and past magnetic field configurations, prevents them from observing directly the dynamo that may exist beneath their feet.
- The complex of motions that theoreticians say must be present if the dynamism of the earth's magnetic field is to be explained has never been detected.

The article with which this *Mosaic* begins was initially to deal with reversals of the earth's magnetic field. These are known to take place from time to time, occasionally being postulated as triggers of such equally mysterious events as mass biological extinctions.

But almost before we got into the subject it became clear that not only were the reversals of polarity still a subject of deep study and little agreement, but the source of the field itself has still to be established.

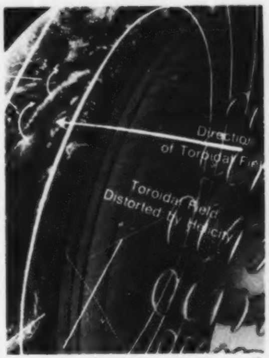
So Derral Mulholland's "When North Becomes South" is not only more than it set out to be but less as well. (That's been proposed as a physical possibility; see "The Coming and Going Star," in this *Mosaic*.) Mulholland's article deals with efforts to define and understand not only the magnetic field's frequent reversals but the field itself. This makes the title a bit of a misnomer, but reversals are still the article's main focus.

Warren Kornberg

MOSAIC

VOL. 13, NO. 5 SEPT./OCT. 1982

NEWARK PUBLIC LIBRARY
NOV 9 1982
U. S. GOVT. PUBS.
DEPOSITORY 376



2

10

24

NATIONAL SCIENCE FOUNDATION

Lewis M. Branscomb, *Chairman*
National Science Board

John B. Slaughter, *Director*
National Science Foundation

Thomas Ubois, *Acting Director*
Office of Government and
Public Programs

Warren Kornberg, Editor
David Heffernan,
Assistant to the Editor
Joyce Latham, Copy Editor

Mosaic is published six times yearly as a source of information for the general public and the scientific and educational communities served by the National Science Foundation. The Director of the National Science Foundation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this agency. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1984. Subscriptions are available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription is $12.00 per year in the United States and possessions and $15.00 all other countries. Single copies are $3.25 domestic and $4.10 foreign. Send check or money order to the U.S. Government Printing Office. Communications other than those concerning paid subscriptions should be addressed to: Editor, *Mosaic*, National Science Foundation, Washington, D.C. 20550. Telephone: 202-357-9776.

Cover. A reserve cut from the Amazon rain forest. (See "Islands of Life in a Forest Sea" on page 24 in this issue.)

R. O. Bierregaard

When North Becomes South *by Derral Mulholland* 2
It is well established that the earth's magnetic field has reversed itself many times. The mechanisms within the earth that trigger the reversals remain a mystery.

From lodestone to dynamo 6

The Coming and Going Star *by William Metz* 10
The bizarre behavior of SS433, an exotic star in the constellation Aquila, tantalizes astrophysicists. A common interpretation of the star's odd activity, related to that of some quasars, is emerging.

Early theories 14

What's in a name 16

Assessing the Risk *by Jane Stein* 17
Many national, corporate, and even personal choices pose inevitable risks. Analyses by risk assessors are helping to balance risk against benefit.

Risks and seat belts 20

Islands of Life In a Forest Sea *by Sam Iker* 24
Deep in the Amazon rain forest, biogeographers are testing a theory of species survival at a time when species survival is becoming a grave problem worldwide.

Superconductivity and Magnetism: An Uneasy Alliance *by Peter Gwynne* 31
The incompatibility of superconductivity and magnetism once was a given. Under certain conditions and in certain materials, marriages can be arranged.

Letters, etc. 37

Women's S/E Doctorates Double Between 1971 and 1981 Back Cover

It is well established that the earth's magnetic field has reversed itself many times. The mechanisms within the earth that trigger the reversals remain a mystery.

Large Scale
Poloidal Field Lines

by Derral Mulholland

y comparison with the earth, the dynamo that produces the sun's magnetic field is both relatively simple and simpler to observe. It is not complicated by the existence of a solid core or an insulating solid crust; since there is no crust, the conducting fluid in which the action is going on can literally be seen by observers.

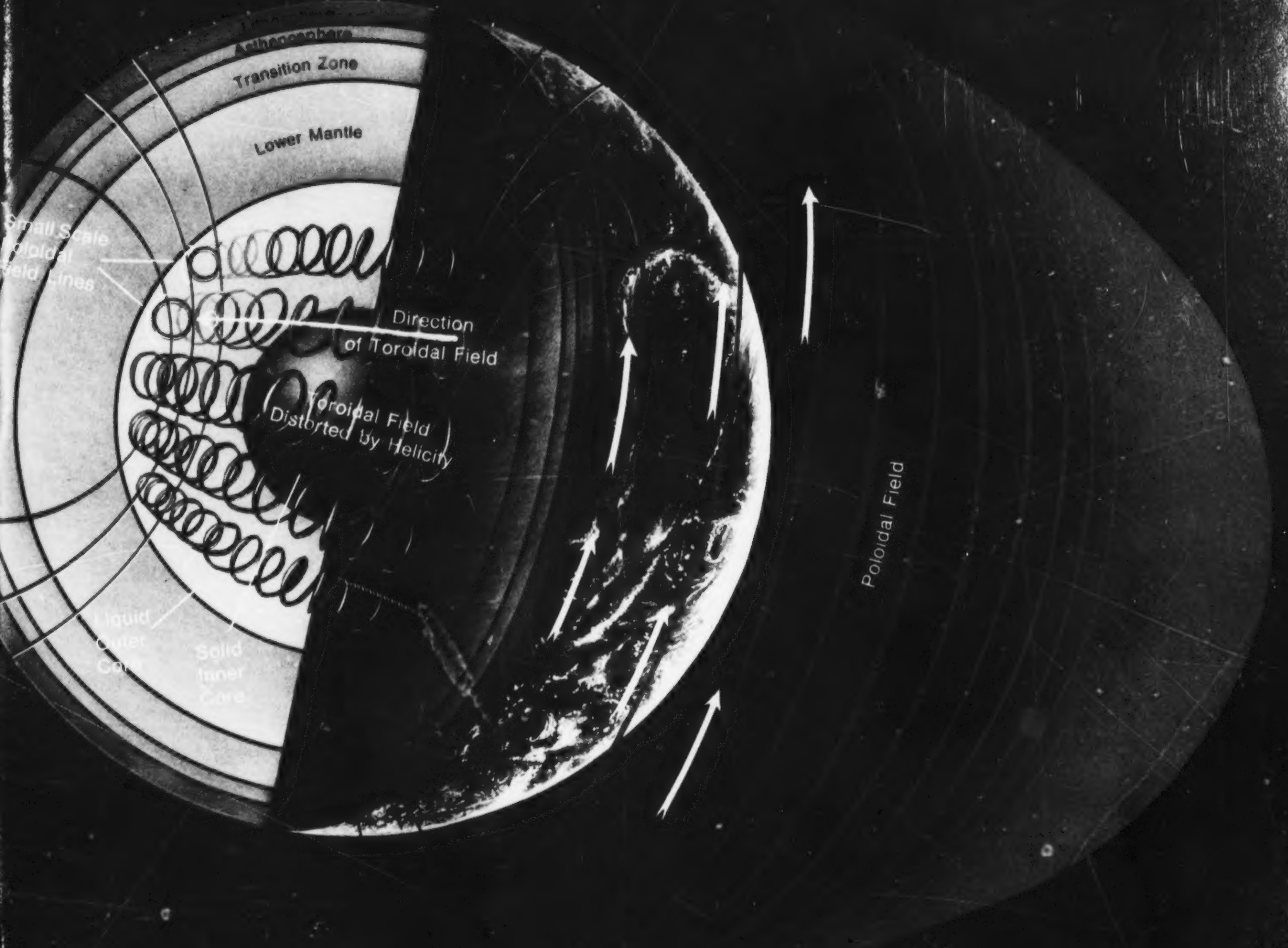
The sun's magnet has a very striking feature. Its polarity reverses quite regularly, every 11 years, at the times of minimum sunspot activity. The reversals almost literally can be watched.

By contrast, the earth's field has never directly been observed to reverse; so slow is the earth's reversal cycle that all of recorded history is not long enough a time. But the evidence is there. That same crust of rock that obscures the physics makes direct observation unnecessary.

The materials that make up the rocks include minerals (such as magnetite) which, having high magnetic susceptibility, align their magnetizations with the earth's magnetic field when cooled to the Curie temperature (580 degrees centigrade), the temperature at which magnetic ordering takes place. Their then-fixed orientations become a record of the field, retained forever in the solid rock. This property is called remanent magnetism. The sampling of it in rocks of different ages is like playing back the earth's magnetic tape recording; it gives a history of the behavior of the magnetic field in past eons.

The fact that the field reverses at all was a chance discovery made by physicist Bernard Brunhes while he was studying the magnetism of rocks from central France in 1906. Since Brunhes's time, the terrestrial records have been played back from many different places over spans of many millions of years. His result has been repeatedly confirmed and amplified. While the earth is not nearly as regular as the sun, the magnetic field appears to have reversed direction on the average of about every million years.

That this is an average needs to be stressed; the rate varies enormously, ranging from periods of tens or hundreds of thousands of years to tens of millions of years over geologic time. The available evidence shows that the reversal frequency hit a minimum of about once every 50

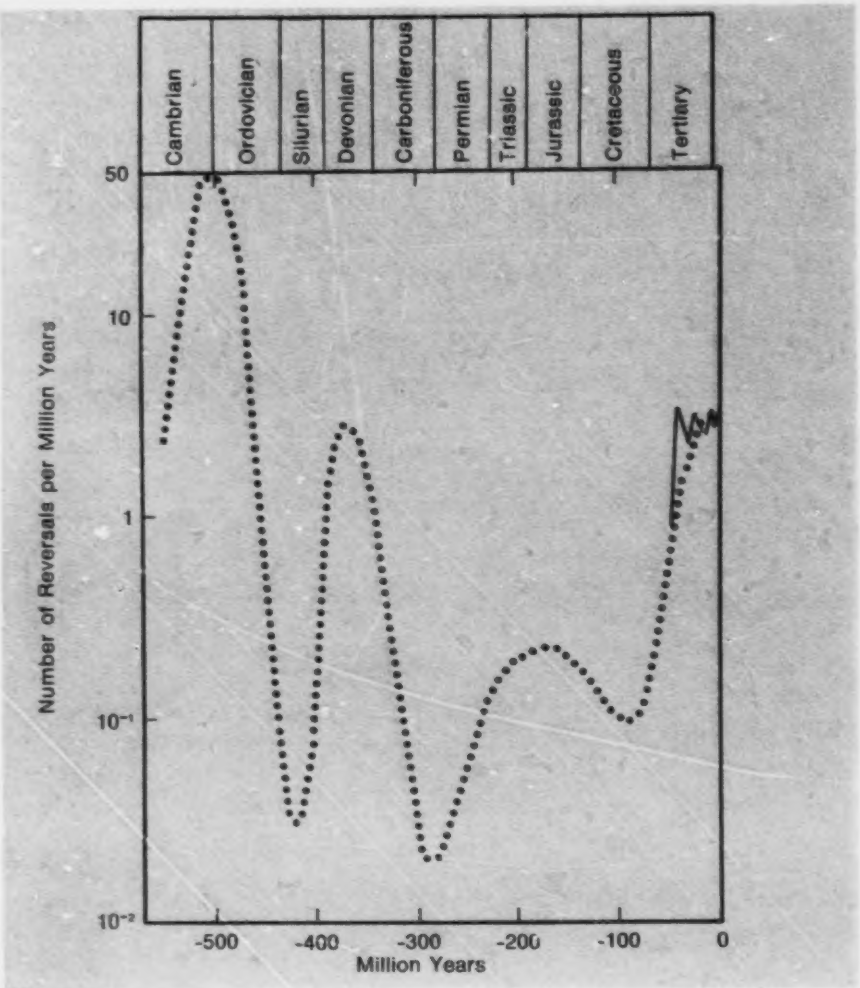

Asthenosphere
Transition Zone
Lower Mantle
Small Scale
oloidal
ield Lines
Direction
of Toroidal Field
Toroidal Field
Distorted by Helicity
Poloidal Field
Liquid
Outer
Core
Solid
Inner
Core

million years at the boundary between the Carboniferous and Permian eras, some 300 million years ago. The field flopped around at the astonishing rate of once every 20,000 years near the end of the Cambrian, 500 million years before the present. Just now, the field seems to be in an intermediate state, with five reversals in the past million years. The most recent may have been only 20,000 years ago.

Many anomalies

How a reversal might look to an observer is complicated by the fact that the field is not the familiar, symmetrical, smooth one produced by a pure dipole such as a bar magnet. A magnetic map of the earth's surface shows many anomalies that, it is known, represent some physical process in the core. But just what process is a matter of debate. Because of the anomalies, however, geophysicists find it useful to discuss magnetic data in terms of what they call a virtual geomagnetic pole. This virtual pole is the place where the north magnetic pole would have been if the field registered in the rocks were that of a perfect dipole. Local magnetic anomalies insure that the poles from different sites will not agree exactly.

The usual picture of a reversal has been built up from many observations of remanent magnetism trapped in the earth's cooling magma and sedimentary rock. Most geophysicists today agree that it looks like this: Occasionally, for reasons yet unknown, the intensity of the geomagnetic dipole begins to diminish and, instead of recovering after a little dip, keeps on going toward zero. For about the first 3,000 years of this event, the dipole is still so much larger than the local field deviations that the decrease does not dramatically affect the location of the virtual magnetic pole.



Reversal rate. In an approximation of data gathered by various investigators, the curve suggests the rate at which geomagnetic reversals fluctuate over time.

After J. A. Jacobs, *The Earth's Core.*

Eventually, however, the dipole effect is reduced so far that the local effects become more important: the field takes on a more complicated geometry, as if there were many dipoles, each dominant in its own place in the complex magnetic topography. Then, after some 3,500 years of such erratic behavior, the earth's dipolelike field seems to regain its intensity, at the expense of the anomalies, reestablishing itself with an orientation nearly opposite the one with which it began. After about 10,000 years, the dipole is back to normal intensity, but its orientation has changed; a reversal is complete.

The apparent pole path, thus, does not represent a true rotation of the dipole axis. It is merely evidence that the strength of the dipole is changing and that the local anomalies are, for the moment, stronger and dominant. From the evidence, it is impossible to tell how the dipole behaves during this period of local dominance, but most specialists believe that the decrease continues smoothly, with a change of sign as the sums of the fields cancel each other and the dipole effectively passes through zero.

If, as can sometimes happen, the apparent pole goes partway only and then comes back, the event is called a geomagnetic excursion. This appears to be a reversal that did not make it. Recently, geophysicists such as Chris Harrison at the University of Miami and California Polytechnic's Kenneth Hoffman have been looking for clues to the reversal mechanism. If completed and failed reversals are basically the same phenomenon then there are just that many more data to help unravel the problem.

A garbled tape

Finding and confirming reversals and excursions in the paleomagnetic record is not always easy. This is partly because the local anomalies insure that the apparent poles recorded in different places will take different paths, even for the same event. Worse still, not all of the crust yields high-fidelity recordings.

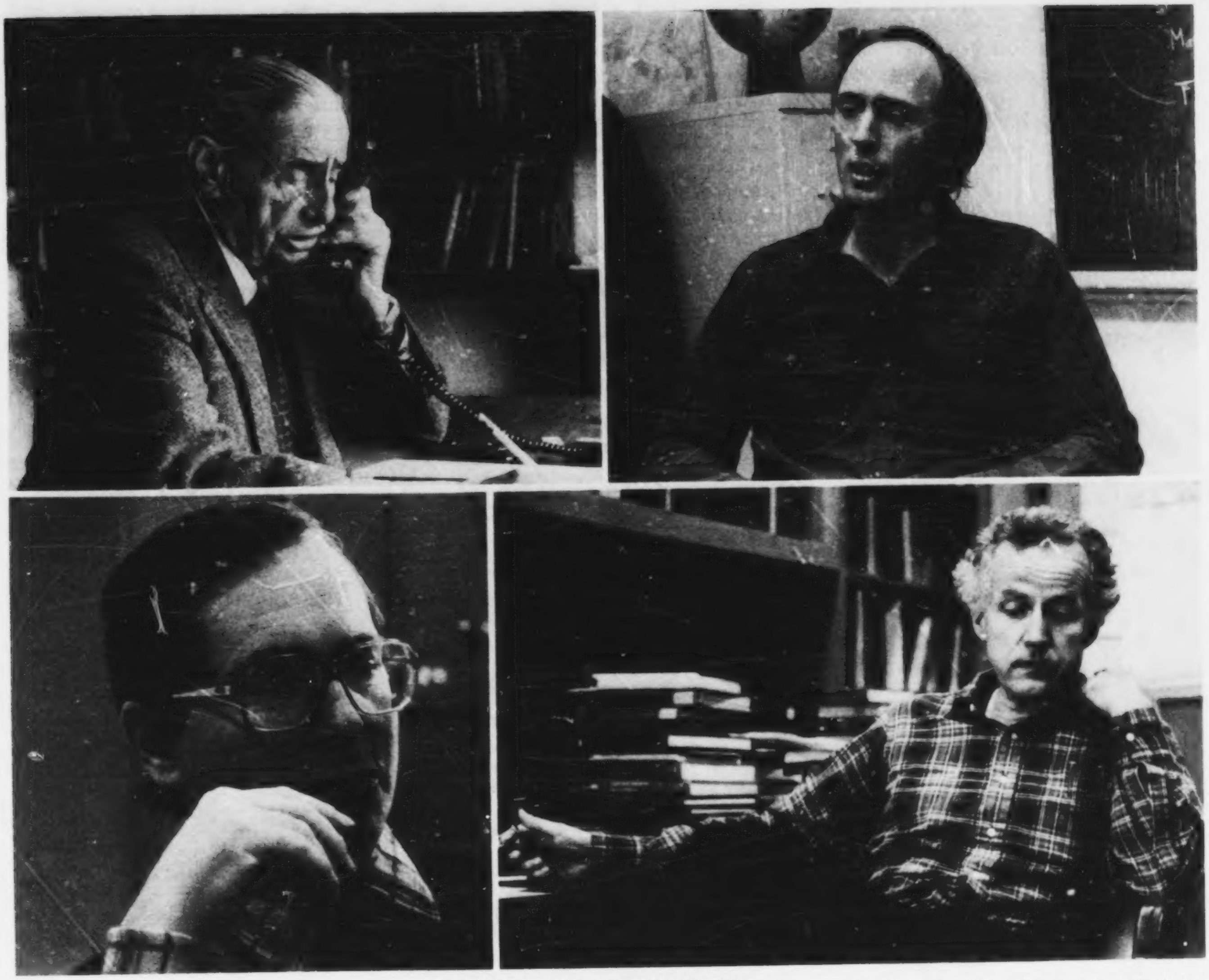
The types of rock found most useful for paleomagnetic studies include lavas, magmatic intrusions of molten rock into fractures in existing solid rock, petrified sediments in old lake and ocean bottoms, and newly created crust. Each of them has unique advantages, but none of them is perfect. Some give continuous records, others accurate intensities, yet others unambiguous directions. None gives all three.

For example, sediments are collected by drilling, and a bad technique will disturb the orientation of the core samples. Variable erosion and sedimentation rates can falsify the record, and the magnetic properties are sometimes disturbed by oxidation. At the University of Minnesota, Subir Banerjee studies lake sediments because he considers ocean sediments to be "hopeless." Miami's Chris Harrison, on the other hand, agrees that some ocean sediments are hopeless, but "other ones are very good recorders, and they are easy to spot."

On land, lavas give fine intensities but are extremely episodic. As Michael Fuller, of the University of California at Santa Barbara, notes, "We don't know if a particular sequence of lava flows all came out in a month or took a few years or a few thousand years." Intrusions of magma from the earth's interior, which cool slowly beneath the earth's surface, provide continuous records of the field, but their magnetization is sometimes hard to interpret, and the field's intensity is sometimes hard to measure.

In the abstract, probably the best paleomagnetic data are those from the phenomenon that revolutionized geophysics by simultaneously satisfying the hypotheses of geomagnetic reversal and continental drift: the magnetic striping of the crust that fans out from the midocean ridges. It is continuous over 180 million years but accurate dating of individual events is difficult, because it has to be done by analogy with continental features. Not only have these not been static, but one of the big problems is a sparseness of Southern Hemisphere data. Calling it unfortunate, Fuller ascribes this problem to the unequal distribution of both land mass and academic centers on the earth.

Mulholland is an astronomer and writer.



Contributors. Walter Elsasser of Johns Hopkins University is the founder of the dynamo theory (clockwise from top left), Peter Olson of Johns Hopkins attempts to construct a realistic physical theory of geomagnetic reversals, George E. Backus of the University of California at San Diego devised a successful geodynamo theory in 1958, and Frederich H. Busse of the University of California at Los Angeles has pioneered inventive laboratory simulations. Derral Mulholland

Preference or parochialism

Judging from the record, the time between polarity changes follows random chance, but the distribution of polarity itself is decidedly nonrandom. The earth prefers a given polarity for long periods of time, although these may be punctuated by brief flips to the other polarity. The present configuration has persisted almost without interruption for nearly 700,000 years, an interval called the Brunhes Normal Epoch. For perfectly egocentric reasons, this period, with magnetic north where it is now, defines what is called normal polarity for the earth. It also defines what we call the north pole of a bar magnet. By extension, the opposite polarity is reverse or south.

In this context, normal means only what is being seen now, not what is most common or natural. Preceding the Brunhes epoch was the Matuyama Reversed Epoch, an interval of some 1.7 million years during which the north magnetic pole was near what is now considered the south geographic pole. This was continuous, excepting about five brief episodes of what would now be called normal polarity. The longest of these lasted perhaps 150,000 years, the shortest 20,000.

How can these reversals be explained? Philosophically, there is no problem. The physical action of a dynamo is such that a particular generating mechanism is capable of producing a magnetic field of either polarity equally well. There is no reason, at least in principle, why an operating dynamo—the accepted geomagnetic mechanism—could not be made to switch poles from time to time.

Mathematically and physically, though, life is harder. First, one must know how the terrestrial dynamo works at all. The problem is not with the dynamo *per se*; engineers know how to make dynamos that work. "The only real problem," says George Backus at the University of California at San Diego, "was whether you can have a single connected lump of fluid that can sustain a dynamo." After that, one can worry about how to make it reverse. Backus was one of the first to prove mathematically that the fluid dynamo could work.

Actually, to a nonspecialist, the issue seems not so much the lack of a solution but rather too many solutions. Every dynamo theorist seems to have an answer, but the answers all differ. One of the reasons is easy to find: While it is possible to write down the general equations that govern the behavior of the fluid core of the earth, they are so complicated that there is almost no hope of ever having a complete and correct solution to them. As geophysicist Friedrich Busse of the University of California at Los Angeles remarks, it is necessary to choose simplifications to make any headway at all. Backus agrees, saying that each investigator makes his particular choices based on what he thinks might lead to useful results: "Everybody has strong ideas about what is reasonable, partly because the field is really not settled yet," he explains.

The mathematical difficulty is compounded severely by a physical one. The geodynamo action occurs in a fluid shell bounded on the inside by a solid conductor, the earth's iron inner core, and on the outside by a solid nonconductor, the crust and mantle. Not only does the crust complicate the mathematics, it also prevents direct observation of the liquid outer core. So pitifully little is known about what is really happening down there.

The choice of simplifying assumptions depends in part on what one imagines to be reasonable fluid motions. Since all that can be seen is the external magnetic field, it is necessary to make educated guesses about the interior, then follow the mathematical consequences to see if they work. In other words, intuition plays a role, and the result is a bewildering multiplicity of models.

Some of them sort of work. Some of them do not work at all. According to Peter Olson at Johns Hopkins University, none of them satisfies the observations well enough to be accepted by everyone working on the subject.

Accepting the necessary

Central to this multiplicity is the fact that every model requires the existence of two different magnetic fields. The external field, the one that looks roughly like a bar mag-

From lodestone to dynamo

In 1906, French physicist Bernard Brunhes was doing field work in the relatively new area of paleomagnetism. The fact that the earth has a magnetic field had been known for more than three centuries, but that information had led to little beyond a rational explanation for how mariner's compasses really worked.

The most primitive idea about the nature of the earth's magnet was a simple analogy to nature. Pieces of magnetite had been used as direction finders by the Chinese since at least 1100 CE, and there is a persistent suggestion of their use as early as 2700 BCE. During long centuries, the force that attracted the needle was unknown. Many supposed that it came from the north pole of the sky, some irresistible heavenly attraction. In 1600, English scientist William Gilbert investigated everything then known about magnetism and concluded that compasses work because the earth itself acts like a huge bar magnet. It was an easy step from there to suppose that the whole planet was a big lodestone, with the magnetism fixed in the rock.

That idea quickly ran into difficulties. The earth's magnetic field changes with time. Following Gilbert's work, careful observation of compasses showed that the position of the earth's apparent bar magnet, which is not aligned exactly with its rotation axis, drifts slowly westward with respect to the fixed crust of the earth.

One of the most inventive minds in the history of science was bent to this puzzle, and Edmund Halley proposed in 1692 that the earth was hollow, with an inner sphere that rotated at a slower rate than the outer shell. It was the inner sphere that was magnetized, hence the drift, said Halley. Considering that conventional wisdom at the time held the earth to be solid, this was a remarkable hypothesis. It was wrong, of course, but it was the best there was until two centuries later, when the new science of seismology would give the first evidence of a fluid core.

The 19th century was a period of intense discovery and invention in the study of electricity and magnetism. The 1820 announcement by Danish physicist Hans Christian Orsted that electric fields influence magnetic fields set off a flurry of activity. Over the next half-century, there was the invention of the electromagnet by France's François Arago, his countryman André-Marie Ampère's hypothesis that magnetic fields are related to microscopic electric currents in the magnetized object, and the near-simultaneous discovery of electromagnetic induction by England's Michael Faraday and the United States's Joseph Henry. This last step gave the basic principle for the electromagnetic dynamo, a reversible machine that can use a magnetic field to create an electric current or use an electric current to create a magnetic field.

Coming together

The earth sciences lagged behind physics for a while, partly due to the persistence of the idea that the planet is no more than 6,000 years old. English geologist Charles Lyell had demonstrated the earth's antiquity and launched the basic division of geologic ages in the 1830s. Since magnetism was first known from rocks, it was a natural thing to measure and compare the magnetism from different sites and ages, a study called paleomagnetism. Most rock bears no magnetism at all, but often tiny grains of magnetite and similar materials are scattered through the structure. These retain very weak traces of the magnetization that was locked into them at the time the rock cooled, recording the strength and orientation of the magnetic field at that time and place.

Volcanic lavas were thought to be good magnetic recorders, and so in 1906 Brunhes had gone off to one of his favorite places, the Massif Central. In this great uplift zone peppered with ancient volcanic domes, covering one-sixth of the French heartland, he gathered his samples carefully for later analysis.

As he did so, he may already have heard that English geologist Richard Dixon Oldham had just declared that the earth's core is molten, inferred from the way the seismic waves from distant earthquakes travel through the planet's interior. If so, it surely was not clear to Brunhes that this had any bearing on his interest in rock magnetism.

Back in Paris, as he studied his results from different places and different epochs, a curious fact drew Brunhes's attention. He had expected some variation, if only from the known westward drift of the dipole (see main story), but he certainly had not expected the variation he found. The rocks showed two basic polarities, each 180 degrees out of phase net, is called the poloidal field, and its main component is the axial dipole. This can be generated only if, within the earth, there is an electric current interacting with a ring-shaped magnetic field: the toroidal field.

Because the crust is an insulator, neither of these two fields can be detected from the surface, but absolutely everyone in the field accepts their existence; there is no way to create the observed dipole without them. What sort of dynamo works inside the earth depends on the strength of the unobservable toroidal field.

On this basis, the first plausible model of the dynamo was proposed in 1955 by University of Chicago astrophysicist Eugene Parker. It was based on a property called helicity, which involves a spiralling motion of the fluid conductor. It is now broadly felt that fluid dynamos cannot work successfully without at least local helicity, but there is a great deal of disagreement about how important a role it plays.

In models in which fluid convection is a significant factor, such as those of Busse and Olson, this helical flow is about the only thing that counts, generating both the toroidal and poloidal fields. In Parker's model, and many others developed since, the helicity produces only the poloidal field while the ring magnet results from strong fluid shear. In the convective models, the two magnetic fields turn out to be of about the same strength, while the shear models result in toroidal fields hundreds of times the strength of the dipole.

Progress and cobbling

There is no question about the existence of fluid convection in the earth's liquid core, but theorists disagree on whether it is important to the functioning of the dynamo. At least one, George Backus, admits that his choosing to ignore convection is a matter of mathematical simplicity.

Until about a decade ago, the fluid-dynamic problem was considered to be so formidable that it was usually avoided. The first step of such a dynamo study—the so-called kinematic approach—was to assume a fluid velocity pattern, without worrying about whether it was physically possible or not. This made the mathematics simple enough so that self-sustaining dynamos could be produced, but the interior of the earth had to be presumed to follow certain physical rules. Thus these studies gave useful information on dynamo behavior but were not directly applicable to the physical problem.

More recently, a more sophisticated and demanding approach has gained ground, pushed by such researchers as Busse and mathematician Willem Malkus of the Massachusetts Institute of Technology. In what they call their physical models, the velocity pattern is required to satisfy the fluid equations of motion. They still cannot use the full equations of the real earth's core, which may always remain too complex for solution, so they continue to search for simplified approximations that will be both realistic and tractable. As Backus remarks, "our mathematical technology is so rudimentary that the solid results have come by cobbling things together with intellectual hairpins and chewing gum."

The wryness of the comment does not detract from the enormous progress in dynamo theory during the past four decades. Still, in a sense, the technical failures of pioneers Walter Elsasser and Edward Bullard (see "From lodestone to dynamo," accompanying this article) can hardly be criticized; their successors have only pushed the boundary at which analysis breaks down closer to the real earth.

In the face of such analytical complexity, it would be comforting to know that a real fluid dynamo can be made to work. Why not build a laboratory model? The combination of high temperatures and pressures that exist in the earth's interior cannot be simulated on an appropriate scale, but maybe some analogous device can be built.

with the other; the earth's magnetic field appeared to have completely reversed itself from time to time in the remote past. A new era in geophysics had opened.

The next clue to the real origin of the geomagnetic field came in 1908, when American astronomer George Ellery Hale stopped organizing observatories long enough to discover that sunspots are centers of intense solar magnetic activity. Half a solar cycle later, in 1919, English physicist Sir Joseph Larmor suggested that the solar magnetic field was caused by a self-exciting fluid dynamo.

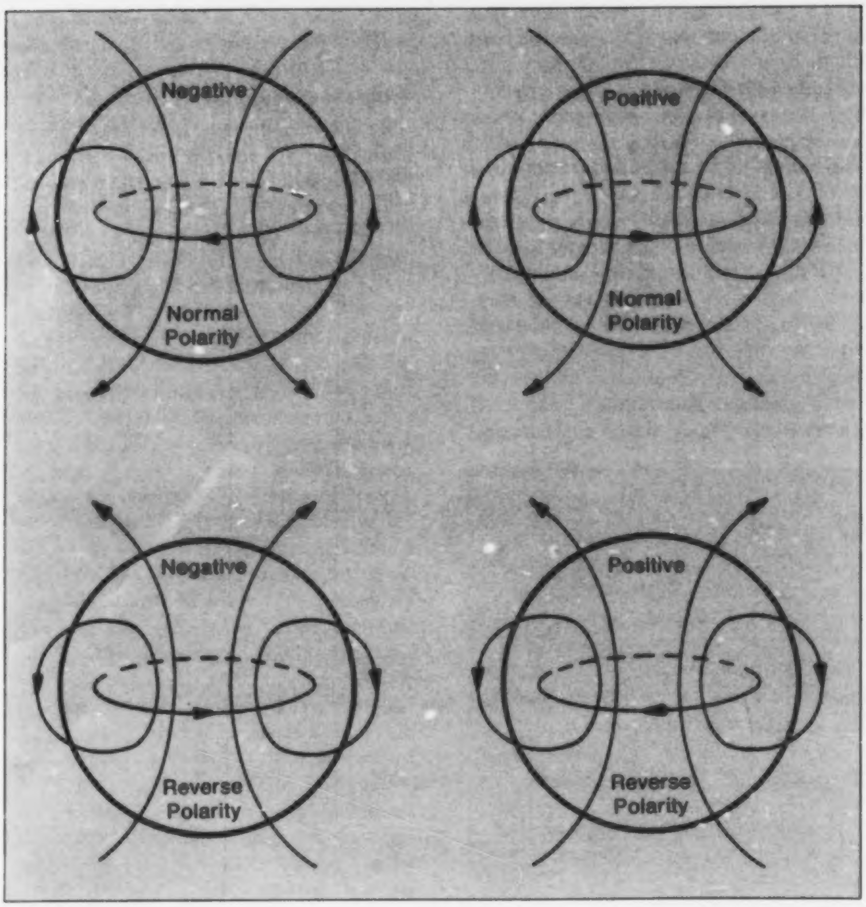
Larmor did not offer details of his dynamo's workings, but his idea opened a new direction for trying to imagine how terrestrial magnetism arises. If Richard Oldham were right, then the liquid iron in the earth's core might produce the same kind of dynamo that Larmor had proposed for the sun. Early attempts to describe the process supposed that there were simple flows and fields inside the earth, all symmetric around the rotation axis. In fact, the mathematical descriptions were so simple that it was possible to prove rigorously that they could not result in a magnetic field. Simple models do not work.

Two decades after Larmor's speculation, German refugee Walter Elsasser came to Robert A. Millikan at the California Institute of Technology, looking for a job. He recalls the sage of Pasadena saying to him, "If you want to stay in atomic physics, I cannot do a thing for you, but if you want to go into geophysics. . . ." Elsasser joined the staff of Caltech as an assistant meteorologist.

The geomagnetic field drew Elsasser's interest. He did not share the general discouragement about the dynamo idea. The fact that simple ideas failed, he argued, did not mean that the concept was wrong, only that theorists would have to work harder. Over the next few years, he wrote several papers purporting to show that a fluid dynamo could be made to work if the fluid flow and the magnetic field lacked rotational symmetry. Independently, English geophysicist Sir Edward Bullard adhered to very similar ideas, following them up in the early 1950s with numerical calculations. In both cases, convection in the core was made to sustain the generation of a magnetic field.

Gradually, the argument gained ground, even though it turns out that there were deadly holes in the proofs. Elsasser's result that a dynamo can work used the right equations but omitted some important physical details. Bullard was computing infinite series on the rudimentary computers of three decades ago; the velocity field that his calculations showed could produce a dynamo action was only an artifact of computer limitations.

It has since been shown that neither man's work really produced a viable dynamo. But that was not really important, because they pointed their colleagues in the right direction—and pushed hard. Now, says Soviet geophysicist S. I. Braginsky, "only the dynamo hypothesis has survived all counterarguments and agrees with the observational data." The frontiers of science can sometimes be pushed in fruitful directions by of incorrect results. •



Schematic. The direction of the earth's magnetic helicity can vary, and with it the polarity of its magnetic field.

Peter Olson

That is probably the case, but it has not been done yet.

Busse, Malkus, and their students have done a number of experiments to investigate aspects of the fluid flows that might exist in the core. Busse has shown that helicity rollers characterizing his theories can be produced in a real fluid. Malkus has simulated the effects on core convection of the precession of the equinoxes, a 26,000-year motion of the earth's orbital plane. Such tests give some confidence in the soundness of basic mathematical approaches, but no one has yet done a realistic experiment, complete with a conducting fluid and a magnetic field.

Busse thinks that he knows how to do it, and he would like to build a test rig. The pinch is that, to simulate the earth's core, he requires several cubic meters of liquid sodium swirling around in a container. Such large quantities now exist only in the cooling systems proposed for fast-breeder nuclear power reactors. Busse believes that the necessary modifications, primarily extra plumbing, could be made to the cooling circuit of such a reactor during its design and testing at reactor research institutions. There are very few institutions operating large cooling circuits, however, and thus far he has had little success in promoting the idea among them.

Pure mathematics

While macrodynamics and other approaches to a realistic physical dynamo theory push onward toward an understanding of the basic dipole, they do not seem to promise any headway on a theory of reversals. To attack this problem, Kay Robbins at the University of Texas at San Antonio has taken exactly the opposite tack, taking the simplest known dynamo model and try to understand it. The model is an electrified rotating disk first devised over a century ago by Michael Faraday and adapted to the geophysical problem by Bullard in the 1950s.

Robbins has no illusions about this dynamo resembling the physics of the earth, but she hopes to learn something about the general nature of the reversal phenomenon. In the 1960s, Robbins notes, MIT's Willem Malkus found that the magnetic field generated by a disk dynamo could be made to reverse by adding an electric shunt, or bypass, to the circuit. Robbins asserts that the physical equivalent to the shunt is the helicity in the fluid-dynamo models.

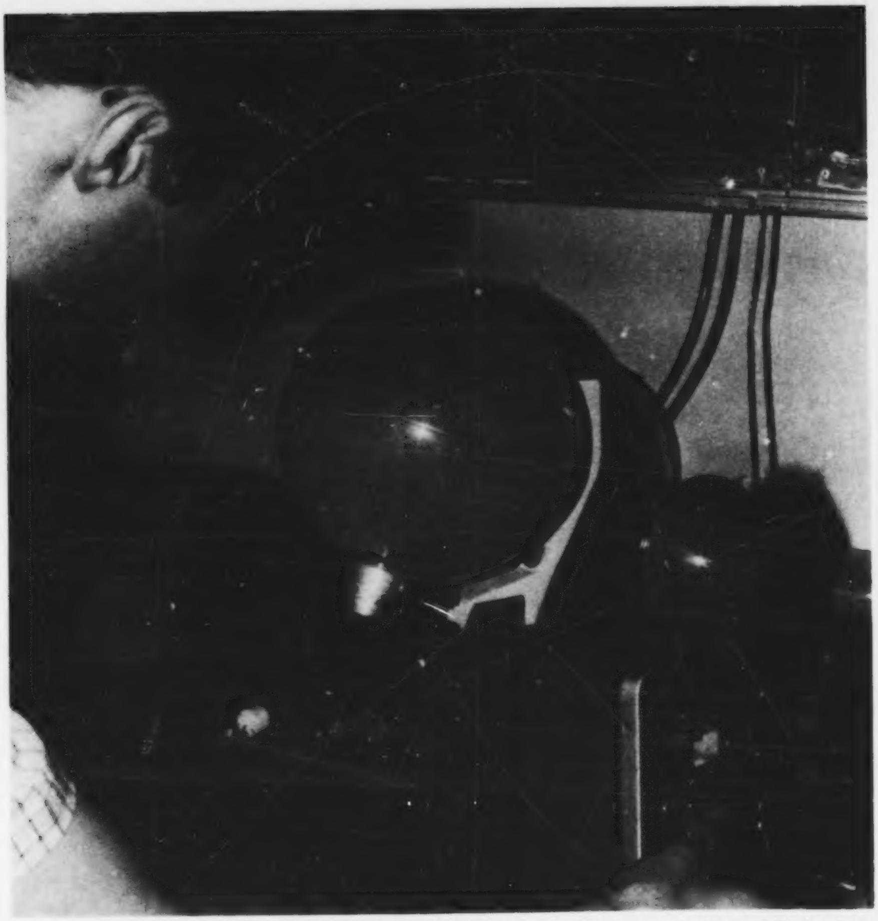
Her technique is largely numerical, with the dynamo equations being solved by an analogue computer, which then shows her the behavior of the dipole solution as a function of time. "This model exhibits varied behavior," notes Robbins. "It will reverse wildly for a while, then it gets stuck in a quiet region. I find a nonuniformity in the reversal frequency that people have been trying to get for a long time." Sometimes, too, there are excursions rather than reversals. Robbins's colleagues consider these results to be very important, but not yet explicable.

Like the earth, some of Robbins's solutions show long periods of preferential polarity. What happens seems to be that the path of the mathematics comes to a fork at which the system must decide to do one or another of two equally acceptable things. Having chosen one fork, it tends to continue down that route, rather than jump back and forth. When occasionally it does jump, it usually stays only a short time before coming back; the mathematical system seems to have a memory and a preferred direction, just as the earth does. The interval between jumps is essentially random, just as it appears to be in the earth.

Since this is numerical mathematics, however, the implication is that the main elements of the reversal process are in the equations, not in the earth. If the reversals are not inherent in terrestrial processes, then geomagnetics as Robbins construes it can never be a predictive science; it will not be known if a reversal is coming until the pole begins to move. Walter Elsasser hopes that this is not so. "I would like to think that there is some physics involved," he says.

The real earth

The strength of the earth's dipole is decreasing now, in what appears to be part of a slow, smooth oscillation. There was a minimum at about 3000 BCE (before the common era), bracketed by maxima around 8000 BCE and 1000 CE, and the dipole strength is headed down again now. Despite these fluctuations, however, there is



Simulation. A machine devised by William V. R. Malkus of the Massachusetts Institute of Technology simulates the effects of precession of the earth's equinox on flow patterns in the fluid core. Graduate student Glenn Ierlie adjusts the device.

Derral Mulholland

no credible evidence for even a major excursion for the past twenty millenia, which encompasses all of recorded history and much more.

There is no reason to suppose that what is now being seen is the beginning of a reversal, but even if it were, it is no cause for panic. Cataclysms are somehow inappropriate to the time scale for a reversal: ten thousand years for the change in intensity, of which the middle thirty-five hundred years see the stately promenade of the north and south magnetic poles to their antipodes.

It was proposed seriously a few years ago that there was a correlation between magnetic reversals and major climatic changes. The implication was that this must be due to the breakdown of the magnetic shield that protects earth from malign influences from interplanetary space. The folklore persists, but the assertions were actually based on spurious data. The observations that were used to support the idea showed no sign of the basic normal and reversed distribution of polarity. According to Chris Harrison, "It was obvious that the samples were not recording the earth's magnetic field, but something else." There is little evidence that the geodynamo affects the weather. (See "The World's Great Dyings," *Mosaic*, Volume 12, Number 2.)

A more serious issue was raised when scientists found that the last major reversal, that from the Matuyama epoch to the Brunhes epoch, coincided with the extinction of a microscopic marine animal, one of the radiolarians. Biologists and geologists agree that the near-simultaneity is real, and it was natural to wonder about cause and effect. Was the reversal the cause of the extinction? Can the failure of the geodynamo cause ecological and genetic disaster?

Calculations were made by D. I. Black of Cambridge University and published in *Earth and Planetary Science Letters* in 1967. He considered all imaginable exterior hazards: cosmic rays, solar flares, dumping of the Van Allen radiation belts, the effect of the solar wind on the atmosphere, and others. In every case the result was the same: Even in the extreme case of ten thousand years at zero magnetic field, a condition much more severe than any known from the paleogeographic record, the direct and indirect hazards to terrestrial life are all negligible.

If you happen to be around during the next reversal, then, it will do funny things to your compass. But it will not frizzle you, alter the course of genetic history, or bring on the next Ice Age. On the other hand, it is not yet known what triggers reversals, and there does remain the possibility that both reversal and extinction can be caused by some other event external to earth.

Many dynamos

The geodynamo is not, of course, an isolated phenomenon. In the solar system alone, at least three other bodies generate their own magnetic fields: the sun, Jupiter, and Saturn. Nor is the sun the only magnetic star.

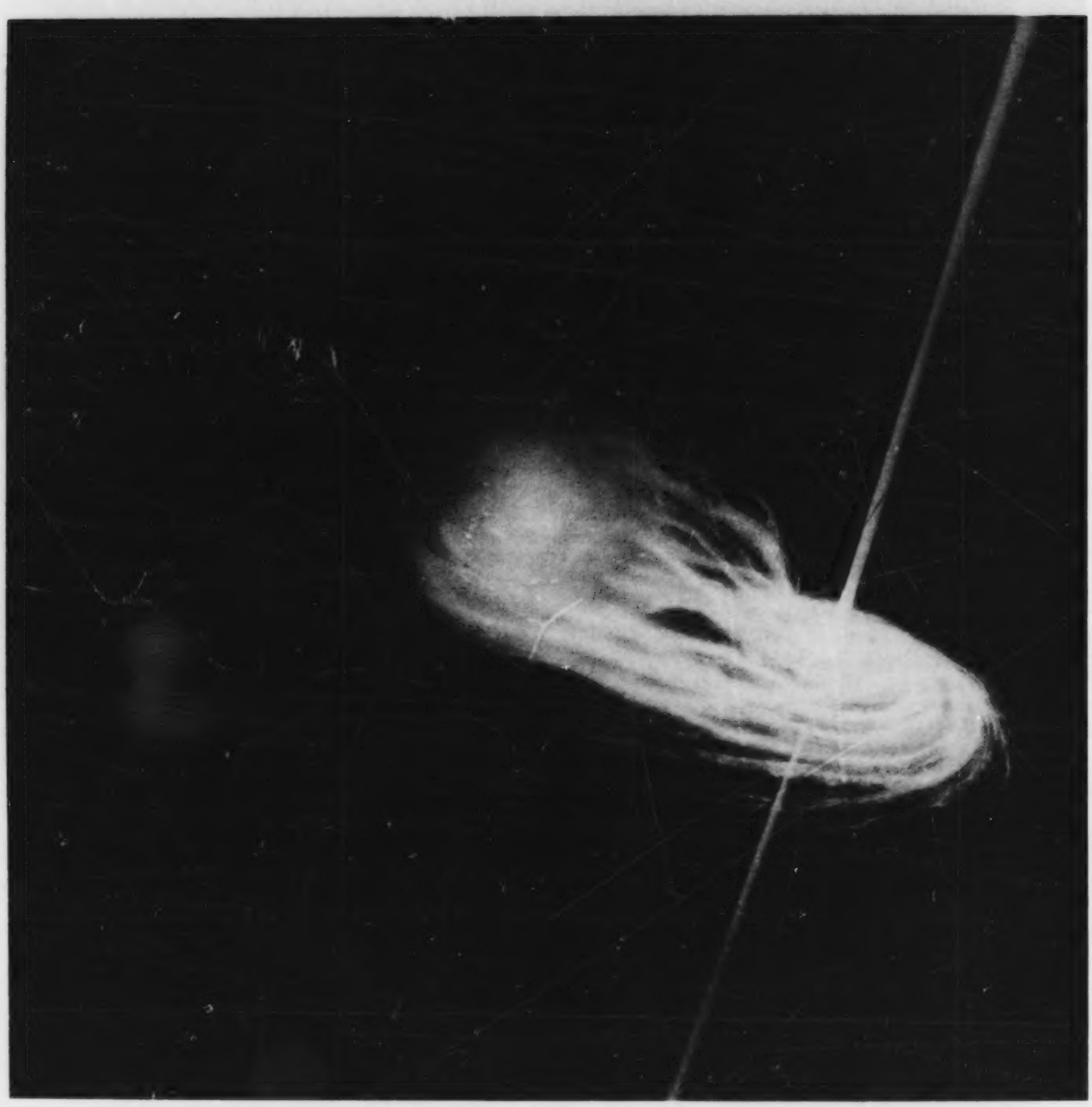
Perhaps even galaxies are dynamos. As Elsasser is fond of remarking, magnetism is a universal fact that happens any time and any place that a conducting fluid is stirred fast enough; the terrestrial dynamo may not even be typical.

Although local circumstances will bring some differences to the various celestial dynamos, each of them nonetheless must operate according to the same general principles. This means that more can be learned about the stars and planets by examining the earth, more about the earth by exploring the stars and planets. Comparison of all their dynamos gives a more ecumenical insight into the physics of all.

The search for a satisfactory explanation for the existence and reversals of the earth's magnetic field goes forward on every scale, from the microscopic examination of fragments of rock to the decoding of signals from the depths of the Milky Way. More direct methods will not be available as long as the core is inaccessible, but the indirect quest spawns a real élan among its practitioners. Says Backus, "I would be disappointed if someone decided to spend a trillion dollars to drill to the center of the earth to find out what is down there. It would take all the fun out of it!" •

The National Science Foundation contributes to the support of the research discussed in this article through its Division of Earth Sciences.

THE COMING AND GOING STAR



The bizzare behavior of SS433, an exotic star in the constellation Aquila, tantalizes astrophysicists. A common interpretation of the star's odd activity, related to that of some quasars, is emerging.

by William Metz

Most stars in the Milky Way stand nearly still in the constellations, moving only imperceptibly over the course of time as measured in human terms. An astronomer can detect these motions with very sensitive equipment, but they are invariably small by astronomical standards. "You never see stars with velocities greater than about 250 kilometers per second," says Bruce Margon at the University of Washington. "There just aren't any."

At the far reaches of the universe, however, some of the most distant objects appear to recede at velocities well in excess of 100,000 kilometers per second. That is one-third the speed of light. And, though there may be surprising velocities among components of distant quasars (see "Faster Than the Speed of Light?" *Mosaic*, Volume 13, Number 4), they just do not occur among the billion or so stars in the Milky Way.

Imagine the astonishment that came with the announcement in April 1979 that a rather bright and normal star just over in the next spiral arm of the sun's home galaxy appeared to be approaching the solar system at 35,000 kilometers per second, while simultaneously retreating from it at 50,000 kilometers per second. It couldn't be doing both, of course, but the numbers alone, not to mention the phenomena that produced them, made the odd star in the constellation Aquila, named SS433, an instant celebrity.

The star appears quite unremarkable in a picture of the constellation. If its light were not obscured by the dust that drifts among the galaxy's arms, it would almost be visible to the naked eye. It not only shines in visible light, but also radiates across virtually the entire electromagnetic spectrum, from X-ray to radio. Through these various emissions it has been watched almost constantly by astronomers around the world, who rushed to start tracking its changing signals as soon as it was reported.

From the wealth of information gathered in the last two and a half years, astronomers have verified what many suspected from the beginning: SS433 is unique among the known stars of the universe—an exotic. Though much has been learned, and broad understanding of the bizarre behavior of the star is growing, important pieces may still be missing.

The common view

Many theories were advanced to explain the odd coming-and-going star (see "Early theories," accompanying this article), especially in the six months after it was discovered. Most of those have now been discarded, and the 12 to 15 groups of theoretical astrophysicists still exploring the phenomenon appear to be reaching a rough consensus. But even "the common interpretation is strange enough in this case," says Margon. Working at the University of California at Los Angeles, he and his colleagues Ronald Downes, Steven Grandi, and Karen Kwitter (now of Williams College), brought the unusual nature of SS433 to light.

The common interpretation is that SS433 is a luminous disc of matter, about the same diameter as the sun, with either a neutron star or a black hole embedded in its center. In this view, SS433 has apparently taken the shape of a disk because it has progressed far beyond the normal state of a middle-aged star and metamorphosed into a nonspherical shape. It appears to be both coming and going because of two needle-thin jets of gas ejected by it in opposite directions and at high velocities. The two jets do not always point at the same spots but trace two small cones in space, each with an opening of 34 degrees. It takes the jets 164 earth days to trace the entire cone.

At any time, one jet points generally toward the earth and another points away. Thus, it is the material being ejected from SS433, not the star itself, which is approaching the earth while receding from it. The angles of orientation that the cones make with the line of sight from earth complicate the details somewhat. But it is ejected material that is the source of the bizarre effect.

The two jets have equal and opposite constant velocities: 80,000 kilometers per second, or 26 percent of the speed of light. As viewed from earth, however, the velocities change because the angles of the beams change in relation to earth. The maximum observed velocities are less than 80,000 kilometers per second because neither beam is ever pointed directly toward earth. The approaching and receding velocities are observed to differ from each other because of the time-dilation effect encompassed by Einstein's theory of special relativity.

Orbiting the disc rather close by, in a binary system, is a faint, normal star from which great quantities of gas are being expelled, to be sucked in by the strong gravitational forces of the disc. As the gases fall into SS433, they generate an enormous release of energy. This is converted into X-rays, visible light, radio emissions, and—of course—the jets.

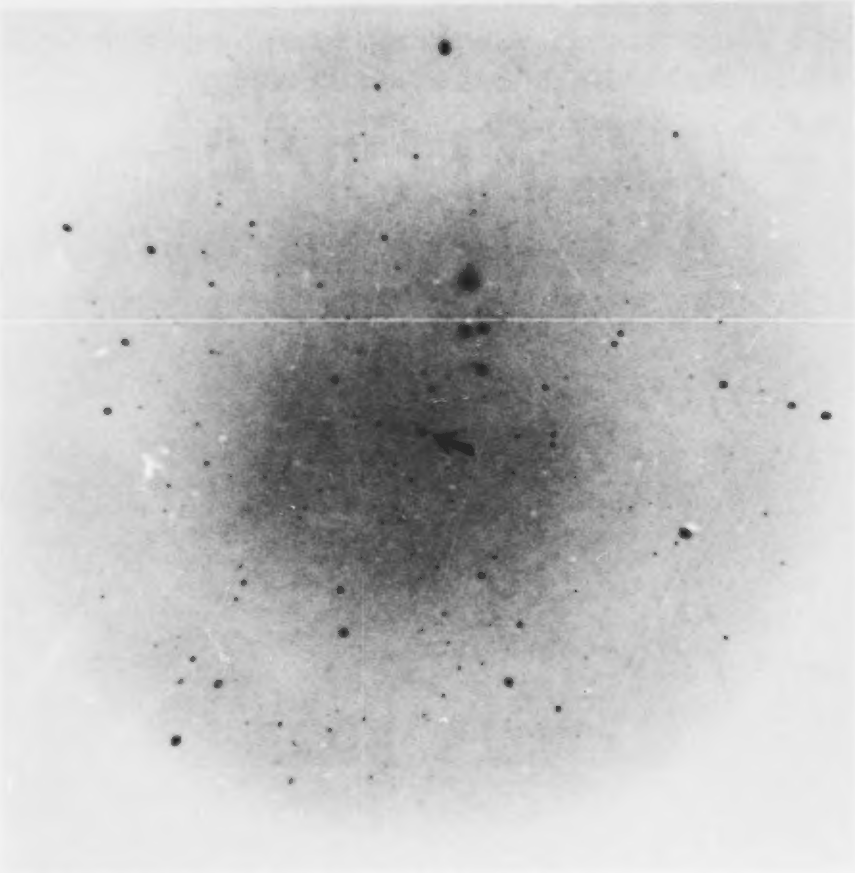
Emerging detail

By astronomical standards, the double beaming of SS433 must be a rather short-lived phenomenon because the jets are ejecting material at a rate that would gobble up the entire mass of the sun in 10,000 to 100,000 years. Since the companion star to SS433 appears to be about the same size as the sun, SS433 must have turned on quite recently in the 13-billion-year history of the universe and is sure to turn off long before most of the stars of the Milky Way blink out.

The jets are quite bright, and it was through the visible light coming from them that the unique character of SS433 was detected. As the two jets trace their cones in the sky, this light is alternately redshifted and blueshifted by the Doppler effect: the change in wavelength that occurs when radiation is emitted by an object that is moving toward (blue) or away from (red) an observer

The exotic. An artist's conception of SS433, based on interpretations of astrophysical data, shows it to be part of a binary system. A compact neutron star (right) is pulling material away from its companion, a normal star. Jets are ejected in opposite directions from the accretion disk surrounding the neutron star.

John Bianchi/*Science Dimension*/National Research Council/Canada



Visible. The optical image of SS433 (arrow) in the constellation Aquila.
University of Washington

Optical astronomers found a periodic pattern to the wavelength shifts of light from SS433. The shifts went from some minimum value to some maximum value and back again every 164 days. It was this pattern of Doppler shifts—once they were accepted for what they are—that made it possible to solve the puzzle of what otherwise appears to be an astrophysical freak—a coming and going star.

Subsequently, radio astronomers were able to confirm the solution that the astronomers working with optical telescopes had found, and they added even more information. Radio observations actually provided pictures of the jets and of the way they interact with the surrounding interstellar gases. The observations told whether the jets rotate clockwise or counterclockwise (they go clockwise as seen from either end). And they showed that matter does not stream evenly out of either side of the disc but gas ejected occasionally and simultaneously from either side.

The one important piece of the puzzle that only optical astronomers could provide, however, was the proof that SS433 is indeed part of a close binary system. Using one of the first 180-centimeter telescopes built in North America, Canadian astronomers David Crampton and John Hutchings detected the faint normal star orbiting SS433. Astronomers had suspected the companion's presence but had been unable to find it. The 65-year-old telescope at the Dominion Astrophysical Observatory in Victoria, British Columbia, was one of the few in the world that was stable enough to detect the normal star (estimated to be of one to two solar masses) and measure its 13.1-day period of rotation about its partner.

Through its various types of emissions, SS433 had in fact been discovered and rediscovered five or six times before astronomers realized its strangeness. Some of these detections gave more accurate positions for the star than others, and by and large the exits in the form of concentrated blobs of various discoverers did not realize that other viewers had seen the star before them. But by late 1978 it became apparent that all these objects were in fact one and the same and that the object in question was peculiar. (See "What's in a name?" accompanying this article.)

Emission lines

The peculiarity that was most immediately apparent in late 1978 was that the visible spectrum of SS433 had strong emission lines in places where they were not expected. Emission lines are strong features in the visible spectrum of the light from some stars. Normal stars do not have them; usually they are found in the spectra of very old or very young stars. They are thought to appear at the odd stages of a star's life when, for some reason, a layer of very hot, ionized gas envelops it. Margon and his coauthors reported in their first paper on the subject that SS433 has bizarre emission lines "which change in intensity, profile, and wavelength daily."

Emission lines usually appear at distinctive places in the spectrum. These places correspond to the excited states of the normal constituent gases of stars, namely hydrogen and helium. Other lines from carbon, nitrogen, sodium, and potassium also show up weakly. But these lines always show up at the same, predictable places in each star's spectrum. (The hydrogen-alpha line is 6563 angstroms, for instance.) That the lines in the spectrum of SS433 moved around from one day to the next was confounding.

The emission lines of SS433 were first seen by David Clark and Paul Murdin, who took a spectrum of the star at the Anglo-Australian telescope in Australia in June of 1978. They knew they had a strange object on their hands because there were not only emission lines but also evidence that the same object was anomalously producing both X-rays and radio waves. Clark and Murdin published their finding in *Nature* and noted that because of the triple coincidence (strong emission lines, X-ray and point-radio source), it was likely that SS433 was something profound.

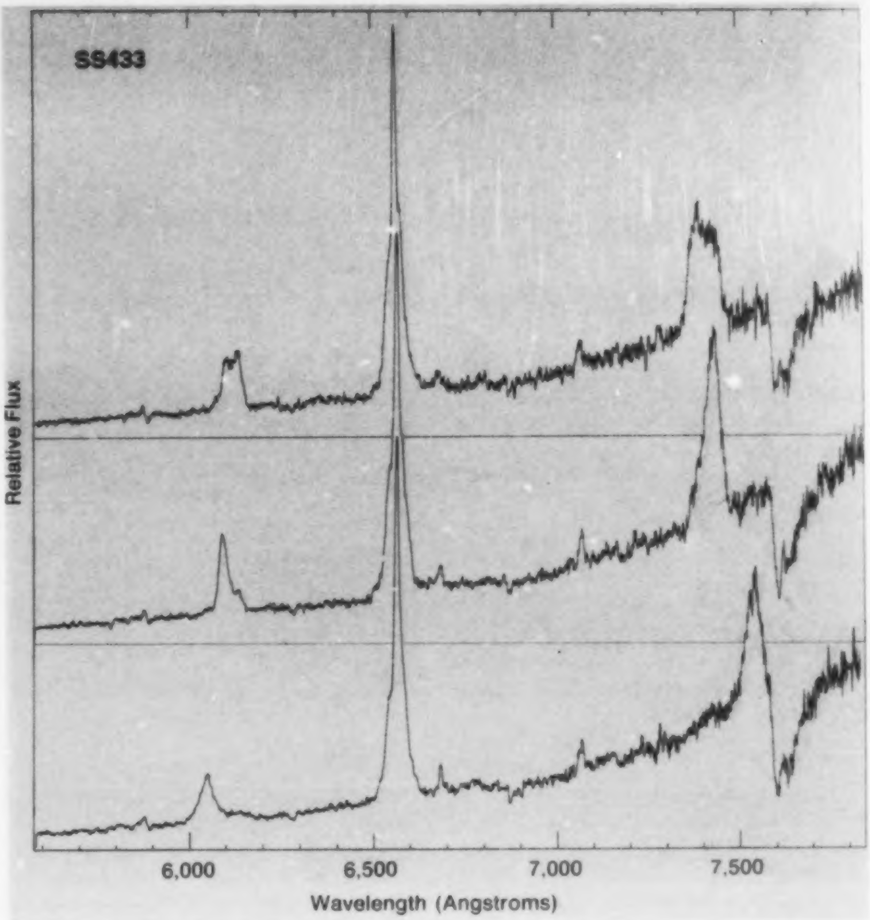
Clark and Murdin were poised to unravel the mystery of SS433 except for one thing. They had just finished their telescope time in Australia and returned to England (which has always been a terrible location for observing the stars and where Clark and Murdin did not have access to a suitably equipped telescope). They were not planning to return to Australia for more than a year.

Metz, a freelance science writer, was formerly on the staff of Science.

Margon happened to be visiting England in the summer of 1978 to work as a guest observer at the only large telescope there, the 250-centimeter Isaac Newton telescope at Herstmonceaux. The weather was so bad he accomplished none of the observing he planned, but he did have a cup of coffee during the night with Clark, who told him about this "funny thing with strong funny lines." Margon soon returned to the United States and within a month was taking spectra of SS433 with the three-meter reflector at the Lick Observatory in California. "I am indebted to Clark for telling me about it," says Margon. "It was the best lost night I ever had."

Within another month, Margon and his coworkers had gathered enough data to publish their first paper—the "bizarre" one—on SS433. Spectra taken on three or four consecutive nights in late October of 1978 were particularly convincing. Not only were there emission lines in strange places, but they were moving in strange ways from night to night.

"One spectrum is enough to tell you that it is something very funny. Two or three spectra are enough to tell you it is something exceedingly funny," says Margon. But then the data flow stopped for a couple of months as SS433 swung behind the sun (as seen from earth) for most of December and all of January. Italian astronomers, working at the Asiago Observatory, had also managed to get a spectrum of SS433 before it disappeared. So the professionally curious stargazers had to mull over the meaning of these strange spectra without any further clues for a few months.



Coming and going. The spectrum of SS433, as recorded on three different nights, shows Doppler-shifted versions of the central emission line (large peak) shifting toward the longer, redder wavelengths (right) while simultaneously moving farther toward the shorter, bluer wavelengths (left). It is this phenomenon that gives SS433 its coming-and-going appearance.

Bruce Margon; Remington P. S. Stone

The puzzle unfolds

One of the first ideas was that the unusually placed emission lines were from some rare element that happened to be grossly overabundant in SS433. According to Ronald Downes of the University of California at Los Angeles, who was working with Margon during this period, "One idea was that the lines were from cesium, perhaps as much as 10^{24} times overabundant. That was Bruce's favorite guess. But it would match up on one night, then on the next night it would not." Italian astronomers looking at SS433 also suggested elements rare in normal stars as the cause. None of these explanations worked for more than one spectrum on one night. They were quickly recognized as inadequate and discarded.

Another idea was that at least some of the lines were artifacts of the observing procedure and not really coming from SS433 at all. Perhaps some of the strange lines were coming from the night sky. (SS433 is so bright that it can be observed optically even when the night sky is not particularly dark.) Observers tried to make the suspected artifacts disappear, but they would not.

Another thought was that the lines were produced by the Zeeman effect of a very strong magnetic field. The Zeeman effect—for the Dutch physicist Pieter Zeeman—produces shifts in the emission lines of hot gases when those are passed through a magnetic field; the stronger the magnetic field, the greater the shift will be. Since neutron stars have very strong magnetic fields, the explanation had a certain plausibility. This idea was entertained quite seriously for a time.

"We knew that there were Zeeman shifts of hydrogen lines in white dwarf stars by hundreds of angstroms," says Margon. Furthermore, neutron stars were known to have stronger magnetic fields than white dwarf stars, so "the thousand-angstrom shifts we saw were not excessive for Zeeman splitting," he adds.

The idea had serious shortcomings, however. White dwarf stars with evidence of Zeeman shifting had absorption lines, not emission lines. White dwarf stars were not generally x-ray sources, and their lines did not drift from day to day. Margon and his colleague, theorist Jonathan Katz (now at Washington University), did think of a way to make Zeeman-shifted lines move in the spectrum, but they soon discarded the idea.

Almost a reflex reaction of astronomers confronted with spectral lines in unusual places is to suspect that they are normal spectral lines (for instance, the lines of hydrogen and helium) displaced by the Doppler effect. It turned out to be the correct explanation, but it was at first rejected out-of-hand by the UCLA astronomers.

"It wasn't that we didn't think it could be a Doppler shift, it was just that we were sure it wasn't," says Margon, reminiscing about that period of mystery and confusion. "Stars were known not to have velocities of 50,000 kilometers per second," he notes.

Downes, who was then working with Margon at Kitt Peak, agrees. "From night to night we noticed that there was a slow drifting of the lines, and we thought it might have been the same lines drifting. But it never struck us that it was the Doppler effect, because the velocity would have been so huge." It was not until the Doppler explanation had been tested and accepted that most of the early models were constructed.

A quasar precedent

The working of SS433 was not the first time a puzzle in astronomy, which turned out to have extremely far-reaching consequences, had turned on the realization that strange emission lines in a stellarlike object were really familiar lines shifted a huge amount by the Doppler effect. "We had talked about Doppler-shift interpretations for a long time," Maarten Schmidt says about quasars (quasi-stellar objects) which he discovered while at Caltech in 1963. "But somehow we were not sufficiently prepared." Schmidt was writing an article on the strange emission lines of the quasar 3C273 when the idea came to him. He says he had discussed possible explanations over the lunch table with his colleagues for about six weeks before he figured it out, and that some of the speculations were much more radical than the explanation turned out to be.

One astronomer suggested that Doppler shifts great enough to convert X rays to visible light might be seen. "When we found out that the redshift of 3C273 was only 16 percent, we were really excited," says Schmidt. The redshift observed in SS433 is in fact greater: 26 percent of the speed of light.

When Margon and his colleagues at UCLA first observed SS433, it had been near its maximum Doppler shift. This was purely coincidence, but it probably slowed down the interpretation of the object. "If we had begun observing as it was east rather than west, we would have observed the entire cycle without interruption," says Margon. But by mid-March, SS433 had reemerged from behind the sun. The astronomers at UCLA had almost a dozen more spectra, and their Italian colleagues had also accumulated more data. The evidence became overwhelming. Margon made a plot of the data from all the known spectra on March 18, 1979. The Doppler explanation fit.

At that point, the UCLA group did not yet have data from the full cycle, but they were pretty sure SS433 was periodic. Margon sent out an International Astronomical Union telegram in early April saying SS433 had a period of 160 days, plus or minus three. "We nearly got it right," he says. By May 9, Margon and George Abell at UCLA had submitted a much more complete account of the Doppler shifts, giving indisputable evidence of red- and blueshifts and properly pinpointing the period of 164 days "based on the observation of parts of three cycles."

In late spring, Margon gave a talk on SS433 that was widely reported. By June 21, the Abell and Margon paper was published in *Nature* and the idea of rotating, relativistic jets was put forward. Many other models were still to be raised and discarded, some more quickly than others.

The observations that have done the most to confirm the rotating jet model are the extensive radio maps of SS433. Many of these maps have been made with a multiple-antenna radio telescope operated in the desert near Socorro, New Mexico, by the National Science Foundation. In many ways, the Very Large Array Telescope, planned and built long before SS433 was discovered, is almost perfect for studying the exotic behavior of the jets. (See "The V[ery] L[arge] A[rray] Turns On," *Mosaic*, Volume 9, Number 2.)

The appearance of the jets depends on the orientation of the observer. They would trace simple patterns if the cones within which they move were pointed toward earth.

Early theories

During 1979, many strange ideas were put forward to explain SS433. Some astronomers rushed into print even before there was more than a sprinkle of data. The most spectacularly successful of these was Mordechai Milgrom at the Weizmann Institute in Israel, who predicted the period and enunciated the rotating jet model, all on four pieces of data. Others put forward detailed countertheories after the kinematic data were available. To achieve the high spectral shifts observed, all the models incorporate some stream or surface that was moving at about one-quarter the speed of light.

Infall model. Since there are two jets aimed in opposite directions, the kinematic data for SS433 would be the same whether matter was ejected out of the jets at 26 percent of the speed of light or was falling in at that speed. Either way, one stream is coming and one is going. The infall model suggested that there were two infalling streams. The problem with this model was its inability to explain what would pour in so much matter at approximately the same rate from top and bottom at the same time.

Exploding shell model. This postulated a shell around the star, moving outward at 26 percent of the speed of light, illuminated from the inside with a flashlight of astronomical proportions revolving every 164 days. Assuming the shell to be transparent and the light visible from earth, the illuminated spot would seem to be coming and going, depending on whether it was on the near side or the far side of the shell at any time. The model fit the optical kinematic data but was ruled out by the radio observations.

Rings-of-Saturn models. Again, a rotating flashlight revolving on a 164-day period is needed, this time to illuminate different portions of a ring of material rotating at relativistic velocities (26 percent of the speed of light). Such a fast-moving ring has to have its constituents kept in orbit by a very strong gravitational force. A small ring could be kept in orbit by a one-solar-mass black hole, but it would be too small to put out as much light as observed. To get a more acceptable rate of optical luminosity, the model would have a 10^8-solar-mass black hole in the center of the ring. Radio data tend to rule out this model too, though some variants have reappeared with features (mass ejected through nodes in the rings) very similar to relativistic jets.

Laser models. The energy loss of SS433 appears to be greater by a factor of ten than that of any known star. The energy loss in question is simply the kinetic energy of matter streaming out of the jets. Certain groups, including some in the People's Republic of China, have postulated that the light from SS433 is generated by a laser process rather than by the more mundane process in common stars. This reduces the mass flow requirement and cuts the presumed energy budget. Problems with this idea arise because—apart from the Doppler shifts—the spectrum of SS433 looks like the spectrum of some of the most common gaseous nebulae, and these presumably do not lase. •

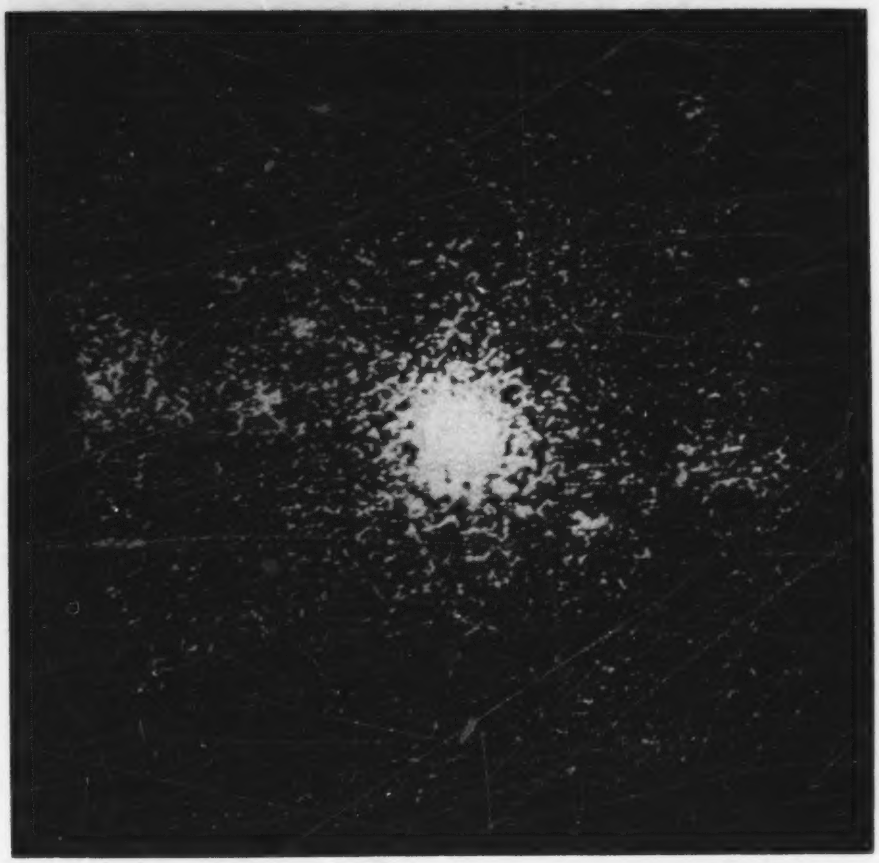
In that case, the jets would trace expanding circles on the sky as blobs of matter spiralled their way out, describing a cone as they went. If the cones were pointed 90 degrees away from the line of sight to earth, the escaping blobs of matter in the jets would trace regular corkscrews as they rotated through the cone.

Corkscrew patterns

Of course, there is no reason to think that SS433 knows or cares whether it is pointed toward earth. It is presumed that the axis of the cone is just pointed randomly; the 79-degree angle that the axis makes with the line of sight from the earth means that the image seen from earth is a not-quite-regular, slightly skewed corkscrew.

Using the kinematic information on the jets, as obtained from optical data, it is possible to map the pattern the jets would describe on the celestial sphere as seen from earth. The expanded skewed-corkscrew patterns change dramatically over the 164-day period of SS433, and the predicted patterns agree quite well with the detailed radio maps made at different times.

It is almost impossible to associate these patterns with rings or shells or infalling gas, as some of the early models suggested, and so the radio data unambiguously point to the ejected twin jet model as the correct interpretation. Out of the complexity of these data, it is possible to extract all the angles in the kinematic model, the handedness of the jets, and even the distance to SS433: 16,000 light-years, or about one-sixth of the way to the far side of the galaxy.



X-ray evidence. A computer-generated photograph of data obtained from the orbiting Einstein X-Ray Observatory suggests the presence of jets (arrows) streaming from both sides of the central object—SS433.

Frederick D. Seward/Smithsonian Astrophysical Observatory.

Some of the most exciting recent radio observations have shown the evolution of the corkscrew patterns over time as material is ejected. Apparently matter is not spewed out uniformly by the jets but in distinct blobs. These blobs always appear in pairs, perhaps as often as every ten days. They are ejected and coast outward, at 26 percent the speed of light, with the huge velocity given by their initial acceleration. Then they continue to glow as radio sources for 30 to 100 days before their radio emissions fade.

"Every time SS433 ejects a new pair of radio blobs it is a new exercise in source evolution," says Robert Hjellming. Along with Ken Johnston, he has studied SS433 closely over the past three years. The most recent result from their New Mexico radio telescope is that one particular pair of radio blobs was sufficiently long lasting to become a double radio source distinctly resolvable from the central radio structure of SS433.

"So we have got a sequence of maps that show a double radio source moving out perceptibly—a few tenths of an arc second every month," says Hjellming. It seems that radio astronomers have not only been able to prove that jets are the correct interpretation of SS433, but that their most powerful telescopes are making it possible to dissect the jets and look at individual pieces—like looking at the beads on a twisted string.

Another striking feature of SS433 is illustrated forcefully, though not exclusively, by the radio observations. That is its symmetry: the structure observed on many distance scales, ranging from the minute to the gigantic, lines up perfectly with the axis of the cone about which the jets precess. The huge radio patch by which SS433 was first noticed, first called W50, is aligned with the axis of the jets. The complex radio patterns now being studied are aligned. Much more minute radio structure measured by long and very long baseline radio techniques is also aligned. And the X-ray observations show an extended pattern to either side of center which is also aligned with the axis of the jets.

The X-ray alignment is apparent from data taken with the Einstein X-Ray Observatory. This dramatic alignment, on many distance scales, is reminiscent of the situation encountered in many radio sources found outside the galaxy; it may be an important clue to the underlying nature of SS433.

Galactic analogy

Astronomers are involved in a science in which it is impossible to do experiments. Thus their first instinct when confronted by something novel is to look for analogies in other parts of the universe. The most immediately obvious analogy is to double radio galaxies, which also have structure that is axially aligned, from the largest to the smallest scales that radio astronomers can observe. Some of these double radio galaxies are also quasars.

For half a decade or more, a number of astronomers have thought that a double radio source was a sort of galactic-scale Roman candle, always shooting out balls of fire in opposite directions. These sources are tens of billions of times as rich as SS433, being whole galaxies rather than single stars. Still, given the limited choice, analogies are used where they can be found.

There is also the deeply felt, but seldom discussed, reason that some astronomers believe the analogy may well be correct. First of all, a favored explanation of the powerhouse that drives double radio sources and quasars is a black hole, with a mass 10^8 times the mass of the sun. (Black holes may have any mass, from 10^{-10} to 10^{10} times the mass of the sun. But the value 10^8 is the one that usually comes out of the calculations for models of quasars.) Some astronomers also believe that the powerhouse driving SS433 is a black hole, with a mass approximating that of the sun. But no one is willing to be committed to the idea because black holes are exotic and, though convincing, remain unproven.

Apart from the gigantic 10^8 leap between them, there may not be any problems at all in making the analogy between jets in SS433 and jets in quasars. "It is a perfect analogy with some classes of quasars," says Martin Rees, head of the Institute of Astronomy at Cambridge, England. He is a theoretician who has developed ideas for explaining a great variety of radio sources and quasars and by any measure is a world-class astrophysicist.

The needle-thin quality of the beams ejected from SS433 is not a problem because there are many radio sources where the jets are no wider than the three-degree angle measured for the coming and going star.

The corkscrew pattern of the jets might also be considered a problem with the analogy. The radio sources earlier identified seemed to have straight-line jets, but many more have now been found with more complicated structures. "Some of these double radio sources have a funny sort of mirror symmetry which suggests that the beams squirting out have been precessing in some way," says Rees. Indeed, some radio galaxies show complex curved structure, and cases have been made for precessing jets.

"The question is, if you scale [SS433] up to a quasar or radio source, would you get objects with slow motion precession?" asks Rees. In keeping with their much greater sizes, the precession periods of quasars would be much longer—up to a million years.

The leap in scale necessary to make the analogy is astonishing, but not a problem. First of all, black holes, the undeclared favorites in the race for an explanation, are about the most easily scalable phenomena ever conceived. Black holes with the mass of the earth, the sun, or ten billion suns would have almost exactly the same physics. Secondly, the change of scale of the flow of matter necessary to fuel SS433 or a quasar may not be a problem either, according to Rees. It all depends on physical factors that track each other.

"So you can argue that if you have quasars and take a miniature version—scale down the mass by some big factor and scale down the accretion rate [of matter] by the same big factor—everything ought to behave the same way," says Rees. "You could have the same sort of flow pattern."

There is a problem—at least in Rees's mind: "It is still not obvious," he declares, "whether the compact object in SS433 is a black hole or a rapidly spinning neutron star." On the quasar scale, the gravitational powerhouse could not possibly be a neutron star; once a neutron star gets slightly more massive than the sun, the gravitational forces become great enough to collapse it into a black hole. So the black hole is the favored explanation for quasars, though this need not be the case on the stellar scale. Nonetheless, and despite Rees's objections that conservatism and the neutron star ought to prevail until proven otherwise, a black hole is probably the secret favorite of most astronomers for SS433 as well.

And so the story comes full circle. Who would have guessed that a presumably nondescript star in the Milky Way would have much in common with the most powerful and distant objects in the universe? In fact, they all looked like nondescript stars at first. Both SS433 and quasars had funny emission lines, and both were found to have huge, unsuspected Doppler shifts. The ways the puzzles were solved had much in common, too, though the discoverers didn't realize the similarity until much later. In the words of Margon, SS433 "may represent a readily accessible small-scale model of these extraordinary extragalactic events."

Put more simply, distant quasars and the coming and going star may be very large and very small versions of the same thing. •

The National Science Foundation contributes to the support of research discussed in this article through its Division of Astronomical Sciences.

What's in a name?

Trappings of the coming and going star, if not the star itself, were first detected in the 1950s. A large patch of radio emission having about twice the angular size of the full moon was found coming from the constellation Aquila. The huge feature was thought, probably erroneously, to be a supernova remnant; it was named W50 for its discoverer, radio astronomer Gart Westerhout, then at the University of Maryland and now scientific director of the U. S. Naval Observatory.

More than a decade later, in 1975, David H. Clark and his colleagues at the University of Sydney made a high-resolution map of W50 and found a bright point radio source roughly in the middle of Westerhout's patch. They didn't know what star (if indeed any star) it corresponded to, but in fact it was the star now named SS433.

The commonly used name is derived from the work of astronomers Bruce Stephenson and Nicholas Sanduleak at Case Western Reserve University in Cleveland. Stephenson and Sanduleak published a list of stars with emission lines found in an area around the constellation Aquila, and the star in question was number 433 on their list. Though the Case Western survey began in the early 1960s, they did not publish until 1977.

At about the same time that Stephenson and Sanduleak published their survey, radio astronomers working with Sir Martin Ryle at Cambridge, England, were preparing another of their famed catalogs of the radio sky. The previous Cambridge catalogs were standard references to what was where for radio astronomers; the fourth Cambridge catalog included still-fainter radio sources which were all designated 4C. They did not know the radio source in question corresponded to anything in Stephenson and Sanduleak's catalog, but they found it in their survey and designated it 4C 04.66.

Hoping to catch some of the cachet of Cambridge, researchers gathering up the great quantities of data from the American X-ray satellite *Uhuru* named their most comprehensive catalog at the time the 4U catalog. They also saw the star in Aquila and named it by its galactic coordinates: 4U 1908 + 05. Astronomers working with the British X-ray satellite *Ariel V* also saw it, to wit: A 1909 + 04. The instruments on these two early X-ray observing satellites were relatively crude. Later, the much better-equipped satellite HEAO-2, also called the Einstein X-Ray Observatory, would pinpoint—but not rename—the bizarre X-ray source in Aquila. •

ASSESSING THE RISK

by Jane Stein

Many national, corporate, and even personal choices pose inevitable risks. Analyses by risk assessors are helping to balance risk against benefit.

As part of the nationwide effort to use new energy sources, the El Paso LNG Company proposed in the mid-1970s to build a terminal on the Texas Gulf Coast to import liquefied natural gas from Algeria. The project promised many benefits: Natural gas is popular for home heating and essential in producing fertilizers and other chemicals. It burns cleanly and is easy to transport as a liquid over great distances. In that form, it takes up less than 0.2 percent of its original volume. Since natural gas is liquid only at very low temperatures—162 degrees below zero centigrade—LNG is not flammable. Commercial markets looked good.

The Algerian project ultimately foundered for a number of reasons. While alive, however, it was a good example of the kinds of assessment processes and techniques becoming available to companies, governments, and even individuals as they consider the risks their decisions may entail.

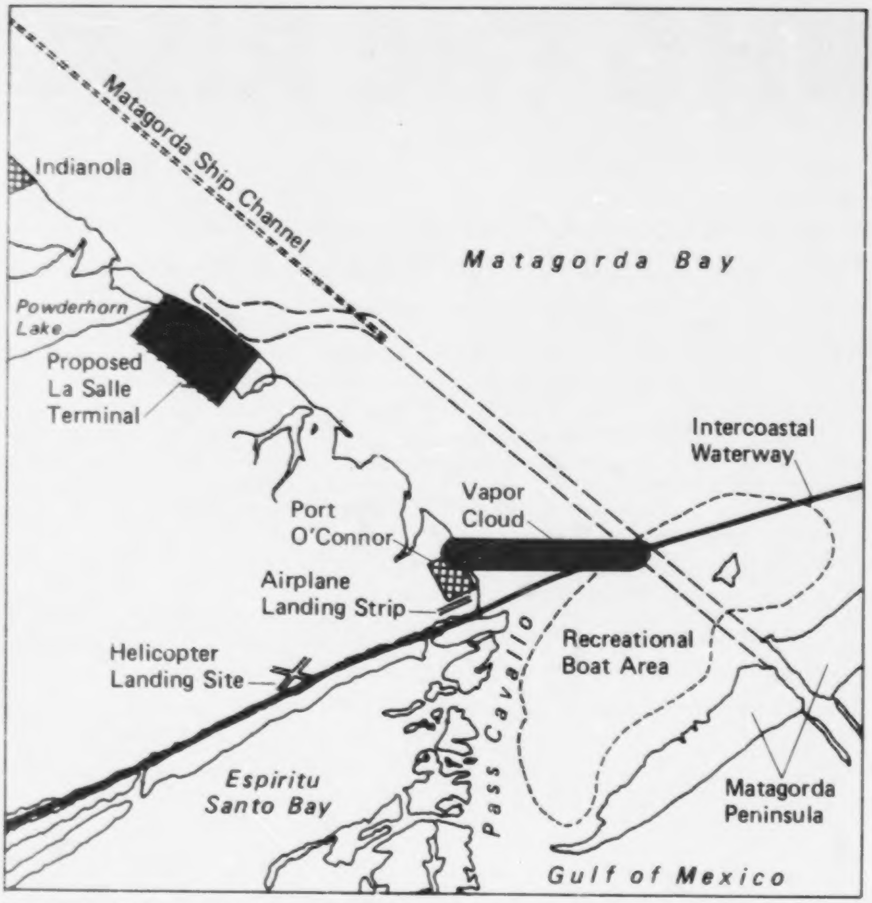
On the risk side of the ledger, if the liquefied gas is exposed to the air, it can vaporize, travel with the wind, ignite, and cause a serious accident. How likely is this? El Paso did not know but wanted to find the answer before committing itself to the costly terminal proposed for Matagorda Bay. So the company did what more and more are doing: It called in a risk-assessment team.

Ralph Keeney and his colleagues at Woodward-Clyde Consultants in San Francisco began by asking: If an accident were to occur, what kind would it be? Based on meteorological and seismic data and the existing safety standards for storing and shipping liquefied natural gas, they concluded that earthquake or hurricane damage to the terminal or to a tanker nearby was extremely unlikely. Among the problems considered worth factoring, however, was the possibility that a small plane from the local Port O'Connor airstrip might crash into a loaded tanker, or that there could be a collision in the harbor.

Copyright material has been
deleted, not available for
Microfiche Edition.

Keeney translated a host of such what-ifs into an event tree—a diagram that shows chains of possible events and the likelihood of each. One possibility was that a liquefied natural gas spill could cause a pool of fire in the harbor; another was that a vapor plume traveling from the spill area to nearby Port O'Connor would ignite over a populated area and then burn back toward the spill area.

The extent of the risk from such a plume would depend on a number of factors: the direction and speed of the wind, the number of people in town (the permanent population of 800 swells to several thousand during the tourist season), the number of



Site proposal. Because a liquid nitrogen gas terminal was proposed at this site in Matagorda Bay, El Paso, Texas, consultants were asked to analyze the possible risks to lives and property. The risks were declared reasonable; the decision not to go ahead with the project was not related to the risk question.

Ralph Keeney/Decision Analysis Group/Woodward-Clyde Consultants.

people on the beach or under the path of the vapor cloud, the number of sources of ignition—from pilot lights to cigarette lighters—and where the plume might ignite.

Following accepted practice in this embryonic and still heavily researched field, Keeney's team assessed for each possible variation the probability of each event in a chain and probable fatalities the chain would produce. The annual risk to an individual life was quantified as two chances in a billion as a result of a liquefied natural gas accident at the Matagorda terminal. Keeney compares that level of risk to the likelihood of a house being struck by one of the 65 meteorites weighing over a pound that land in the United States each year—a far smaller risk than that encountered in transporting equivalent amounts of other forms of energy.

Stein specializes in issues of science and health policy.

Based on Keeney's assessment, the level of public risks at the planned terminal appeared reasonable; on a risk-assessment basis, the project could have proceeded.

What Keeney did not assess—indeed, researchers are still grappling with ways to do it—is the difference between an average or statistical risk and a real catastrophe that might occur despite even the most successful risk-analysis program.

Other risks

Accidents are only one of many risk factors that must be considered before El Paso or, for that matter, any company or government builds a plant, promotes a new product, or embarks on a new venture. Costs, natural and political environments, logistics, and corporate or national priorities are also parts of any such decision. But particularly for decisions involving the impact of new or unfamiliar technology, assessing the extent of risk is becoming increasingly important. "Companies now want to know if they are buying another Love Canal," says Michael J. Murphy of Risk Science International, a Washington, D.C. firm that also advises on environmental liability.

Consider the hypothetical case of a manufacturer thinking of buying a site that has chemical wastes buried on it. What are the risks for the manufacturer? This is one kind of question that firms like Risk Science International are in business to address. After making a site visit, the risk assessors determine what kind of chemicals were used, spilled, or buried and how mobile and toxic they are. Soil and ground-water samples are analyzed to determine how far the wastes have spread. The ground itself is studied to see if the site is on solid rock or permeable ground—factors that could affect the extent of damage (or risk) from the buried wastes. There are also social questions to be answered: How large is the nearby population? Did the previous owner comply with or ignore environmental regulations?—and the like.

In the process of answering such questions, Murphy and other practicing risk assessors can tap into the base of statistical and analytical research that both university and industry researchers are developing. There is also a wealth of pragmatic information that individual consultants are amassing through experience. But because the field is still so new and because the risks themselves are inherently so difficult to analyze, the nature of the information risk assessors get is not always precise.

Companies like Risk Science International have to come up with some solutions anyway, often by quantifying and incorporating necessary imprecision. "We are looking at a real-life situation," says Murphy. "We have to determine whether there is a large enough exposure—or risk—to assist companies in deciding on whether or not they may need additional insurance." Risk assessment, for Murphy, can be a double-edged sword: "It can identify potential human health effects associated with exposure more clearly, but identifying those effects can also increase a company's liability." Nonetheless, he says, it is better to know your own problems first.

There are other kinds of risks. Investment risks in any of 48 countries are the concern of the hundreds of banks, corporations, and governments that use the services of California-based Business Environment Risk Information. This firm's triennial reports tell subscribers what the potential is for a political upheaval, how restrictive the tax structure is for foreign investors, what the chances are that a business agreement made today will be honored a year from now, how available and well-trained the labor pool is, or whether telephone and transportation services are reliable. In short, the reports tell investors things they ought to know before they risk their money.

Using computerized data and expert analyses, BERI advisors make a range of recommendations. They might find that a country is too risky for a long-term investment but safe for deals involving one transaction at a time. But learning that a country is risky is not always a deterrent to investors. "Most places where drilling for oil goes on are risky," says BERI research vice-president Mary McCarthy. "But oil companies go anyway because of the profit potential. They would not take that kind of risk to sell Hula Hoops there."

BERI suggests only whether or not a country is "investment quality; we give [clients] a macro view," says McCarthy. "They have to make the decision of what to do." Most risk assessors agree that that is their role, whatever the subject of the assessment. Joseph Rodricks of Environ Corporation, a Washington, D.C., firm that advises on the risks of chemicals, says: "We assess what the risks are [for a given chemical], not whether they should be reduced or what the level should be in the environment." Indeed, the goal of risk assessment is not even to give an answer. Rather, according to Keeney and others, it is to provide insight and interpret data for decision makers who have to resolve a problem.

That might sound like passing the buck. But the practitioners of this new and burgeoning field would not have it any other way because the risks they are assessing will never go away. Risks are negative consequences that have some probability of occurrence. They can be estimated, controlled, minimized, or managed within some conscionable margins. But they cannot be eliminated.

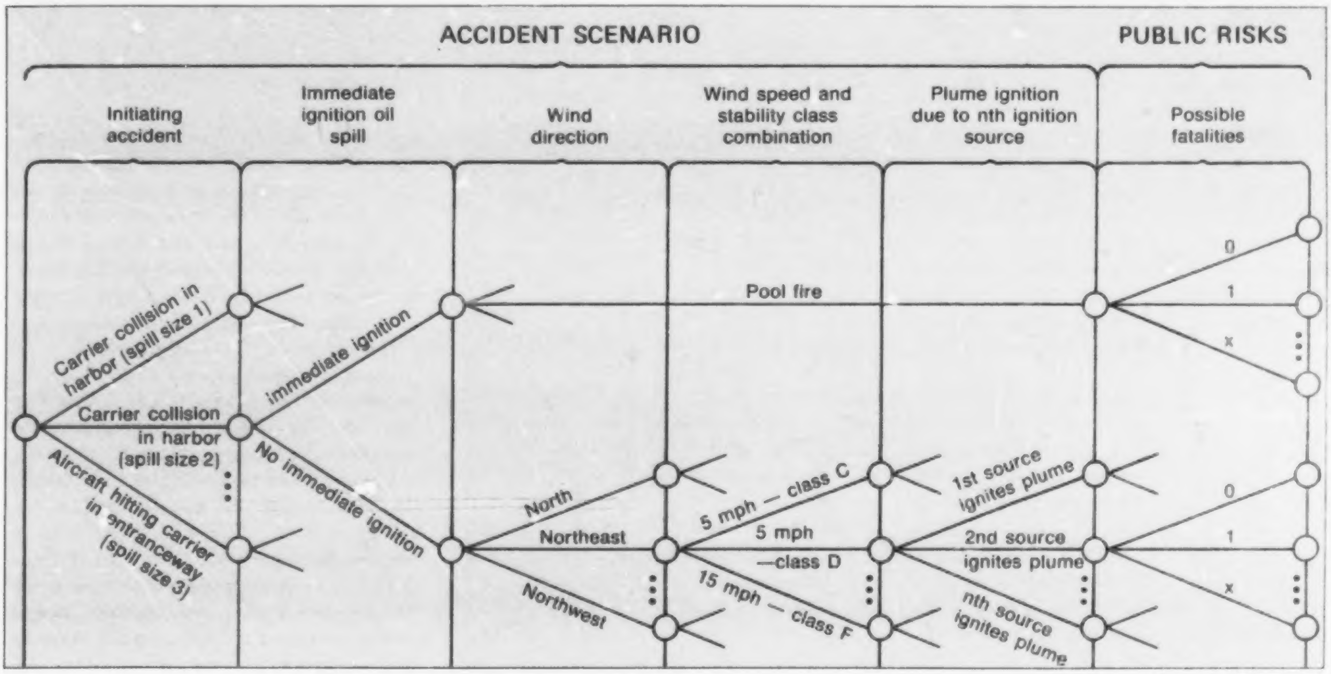
Investigating risk

Living with risk is not a new human phenomenon. However, what is new, or at least increasingly evident, is the extent to which the risks of disaster of one kind or another are the results of human activity—perhaps a residue or unforeseen effect of a valuable activity. Some of these risks are difficult or impossible to predict. But more and more it is becoming possible to detect minute amounts of potentially harmful substances. Along with the advances in measuring these substances comes the responsibility to determine whether they are harmful and at what level of exposure they become a hazard.

Because of these uncertainties, the public is becoming more concerned about—even leery of—the risks in their lives. According to a recent Louis Harris poll, people want more accountability from decisionmakers concerning risks. Ironically, even when decisions are made public, confidence in them is low. One of the reasons for this lack of confidence is often conflicting information. Is saccharin safe to use? Will it cause cancer in people as well as rats? The answer depends on which report you believe.

There is a growing body of interdisciplinary-research to deal with the problems and inconsistencies involved in the assessment of risk. Psychologists are teaming up with economists, management analysts, engineers, political scientists, chemists, and biologists. The field, which has been evolving since the mid 1960s, is still half science, half art. Perhaps it will always be that way. "Who decides what constitutes a risk, what substance to test, or what data to use in analyzing it?" asks Ralph Keeney. Risk assessment, he says "cannot be one hundred percent scientific. It is not objective."

Miley W. Merkhofer, decision analyst at SRI International, a California center for problem-solving research, agrees with Keeney. But, he says, risk assessment is becoming more scientific. "All the information necessary for a decision gets put into a framework. This gives the decisionmaker ways of looking at the problem objectively," he notes. One example Merkhofer has in mind is a procedure he and his col-


ACCIDENT SCENARIO
PUBLIC RISKS
Initiating accident
Immediate ignition oil spill
Wind direction
Wind speed and stability class combination
Plume ignition due to nth ignition source
Possible fatalities
Carrier collision in harbor (spill size 1)
Carrier collision in harbor (spill size 2)
Aircraft hitting carrier in entranceway (spill size 3)
Immediate ignition
No immediate ignition
Pool fire
North
Northeast
Northwest
5 mph — class C
5 mph —class D
15 mph — class F
1st source ignites plume
2nd source ignites plume
nth source ignites plume
0
1
x
0
1
x

Event tree. An accident scenario for the proposed liquid nitrogen gas terminal at Matagorda Bay puts together events and quantifies possible fatalities. The symbols in the public-risk column indicate probability values for fatalities.

Ralph Keeney/Decision Analysis Group/Woodward-Clyde Consultants.

leagues recently designed to assess the health risks of various air pollution standards.

The Environmental Protection Agency was in the process of establishing national ambient air quality standards to protect public health, as required by the Clean Air Act of 1970. Before health risks could be characterized, however, some methodology had to be developed to estimate how serious the risks are and how likely they are to occur.

Assessing air quality

The SRI team's assignment was to find a way to quantify these risks. The team members started out with what was known and essentially undisputed: Best available information was the basis for a computer model of the mathematical relationships between alternative air quality standards and a measure of overall seriousness of the resulting health effects. Also included were statistical uncertainties such as the changes in concentrations of pollutants because of weather.

The first major problem SRI faced was how to factor into the computer model processes that are not clearly understood. Unknown, for example, are how rapidly sulfur dioxide emissions are converted through atmospheric chemical reactions to various sulfate compounds, at what level sulfates cause an increased incidence of asthma attacks, and the effect of carbon monoxide on a driver's reaction time.

Through interviews with health experts, the SRI team obtained judgmental probability estimates for key uncertainties, such as whether a 1,000-megawatt plant in Appalachia burning 3 percent sulfur fuel would raise the annual average concentration of suspended sulfates in New York City by at least one microgram per cubic meter, or whether asthmatics might experience more than a 10 percent increase in asthma attacks for each microgram per cubic meter increase in sulfate concentration. Using the computer model, those probability estimates were converted into curves showing the probabilities of various levels of effects on people's health.

In the final stage of the health risk assessment, dollar values were assigned to health effects to obtain an overall measure of the health cost associated with a given air quality standard. The decrease that could be obtained in this health cost through resolving the various unanswered health questions was then computed.

According to Merkhofer, this is a particularly important exercise because the value of having "perfect information" can be related to the cost of gathering it. As a result of this, decisionmakers can know precisely where they should be devoting their research efforts.

By using this method, Merkhofer explains, an administrator who ultimately must decide what level of risk is acceptable in setting air quality standards can be given clear and concise measures of health risks. These guidelines will help the administrator to discriminate between good and bad proposed standards. The SRI theoretical model is being tested against standards for one pollutant: carbon monoxide.

Limits and trade-offs

As a tool to decide the level at which air pollution standards should be set, risk assessment still has its limitations. The Clean Air Act says that national standards should protect the health of the most sensitive people in the population, with an adequate margin of safety. These are vague terms, Merkhofer points out, yet risk assessors have to translate them into something "precise enough to quantify."

In spite of these shortcomings, Merkhofer argues, the importance of risk assessment as a decision-making tool should not be minimized. "Risk assessment is not homogenous," he says. "There are a variety of fundamentally different approaches, and many criticisms do not really apply to all approaches."

To find out more about risk-assessment tools and what works—and does not work—SRI International is comparing the strengths and weaknesses of three different approach-

Risks and seat belts

Each year more than 35,000 people are killed in automobile accidents. Fatalities could be reduced substantially if more people wore seat belts. But research shows that only about 15 percent of front-seat occupants routinely wear them.

While this is disturbing, it is not surprising to psychologist Paul Slovic of Decision Research. "People respond to hazards they perceive," he says. The problem, according to Slovic, "is that most people properly recognize that the chance of being killed or injured on any particular trip is remote. After all, only about one fatal accident occurs per four million trips, and only one serious injury occurs per 100,000 trips. Therefore," he adds, "you really cannot call people unreasonable because they don't wear a seat belt on every single car trip."

But if these risks were viewed from the perspective of a lifetime of driving, Slovic is convinced that at least some people would change their minds. To test this theory, he and Norman Schwalm helped design a series of pilot radio and television spots for the Department of Transportation's National Highway Traffic Safety Administration. One of them went like this:

Announcer: (a wry, knowing voice of reason.) Oho! You Americans! You do a lot of driving, getting in your car 2, 3, 4 times—call it 2½ trips—a day. 365 days a year. Say 50 years of driving. Do you realize that in your lifetime each of you makes more than 45,000 trips by car? And even though each trip is relatively safe, over a whole lifetime of driving, one out of three of you will be injured seriously in an automobile accident. Those are the odds—one out of three.

Young man: Forget it!

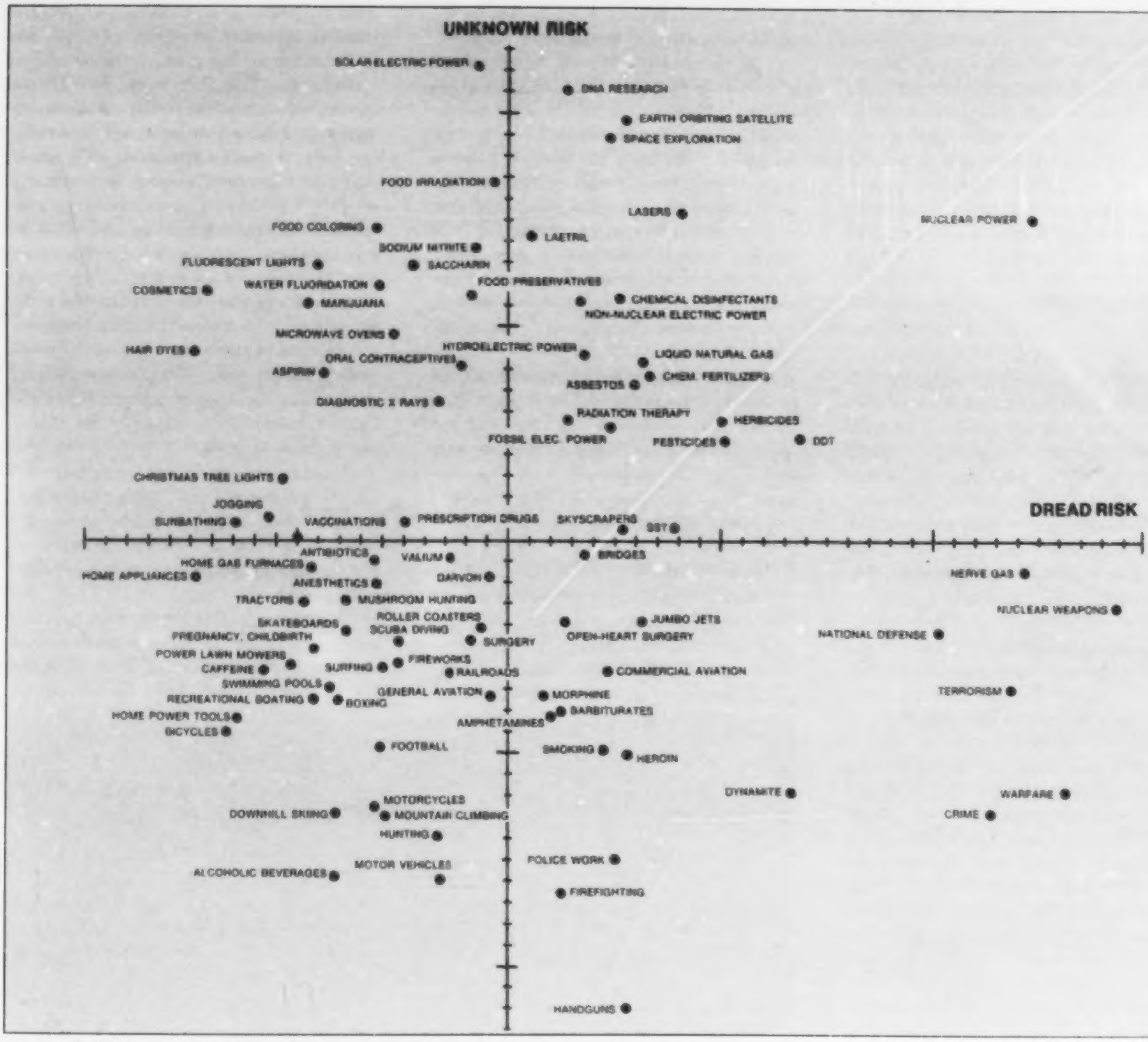
Announcer: Uh-huh! You are all pretty good drivers. Each time you get in a car, you figure you are in control. And most of the time you are right. But remember, you have got no control over those other drivers out there. Or over lousy driving conditions. And when the odds say your time is up, it may not even be your fault, but you will still get hurt—or worse...

Young woman: Look, there must be something we can do to avoid this.

Announcer: Actually, there is something you can do, because none of you is wearing a seat belt. You can make one big decision to buckle up on every car trip, so you are ready when and if the odds finger you. . . . Saving a lifetime means buckling up for a lifetime—starting now.

When this and related messages were presented to a group of 250 people, observed seat-belt use doubled over a month's time—from 15 percent to 30. Among a smaller group of people who heard the message four times a week for a month, the reaction was even more dramatic; as many as 55 percent of the group said they buckled up more often.

"Our research shows that risk-perception messages work," says Slovic. But he is concerned that there is a limit to the numbers that will respond to them. He does not know what that limit is, but Decision Research is trying to find the answer. •



Judgments. Technologies and items were rated on characteristics that analysis showed could be represented by the two factors seen in the grid: dread and how well understood the risks are perceived to be.

Paul Slovic/Decision Research

es: social choice, cost-benefit analysis, and decision analysis.

One way to find out how people think about a risk is through the results of elections or public hearings. A drawback to that so-called social-choice approach: There is no assurance that the popular vote will come up with a response that is necessarily the least risky. For example, several communities voted for the right to use the purported anticancer material called laetrile even though medical research showed it was useless as well as potentially harmful.

Another method, cost-benefit analysis, calculates the total dollar value of alternative solutions to a problem without regard to who pays the costs and who receives the benefits. For example, the cost of accepting 30 days of downtime as the consequence of an accident of assessed probability can be weighed against the cost of reducing that probability. Although cost-benefit analysis seems to be an objective representation of area of measuring dollars against the lives of human beings.

As an approach to societal problems, decision analysis is perhaps the most logical alternative. It recognizes the inherent subjectivity of decision making and breaks down complex issues into components so that uncertainties lie within the accepted domains of recognized experts. Probability judgments and the extent to which consequences are counted as positive or negative are made explicit, and estimates of the likelihood that the consequences will develop are computed. Decision analysis tends to work best where there is one person making a decision—for example, whether or not to develop a new product or build a liquefied natural gas terminal. It does not present a broad public consensus.

SRI is in the beginning stages of developing its risk-assessment taxonomy. Perhaps it will turn out that the best risk assessment incorporates a little bit of each method—social choice, cost-benefit, and decision analysis. But no matter how it turns out, SRI's Miley Merkhofer hopes there will be less confusion about the merits of risk assessment than there is today.

The biggest problem facing risk assessment of new technologies might not be figuring out which methodology works best. Instead, it might be trying to understand how people perceive risks. As psychologists and economists are finding out, risks cannot be analyzed only in terms of mortality, cost-benefit or of many other measurable quantities.

The subjective side

Psychologist Paul Slovic and his colleagues Baruch Fischhoff and Sarah Lichtenstein of Eugene, Oregon, have been studying the way people make decisions when risks are involved. Their research group, appropriately called Decision Research, has found that all forms of risk assessment include a large component of subjective judgment. Often, these subjective judgments reflect errors in perception. When people are asked to judge the number of deaths per year from different causes, they tend to overestimate rare causes of death and underestimate common causes. In addition to this general bias, specific biases are often evident. For example, accidents are judged to cause as many deaths as disease, which actually takes about 15 times as many lives.

In spite of such common misperceptions, most people are quite confident of their assessments. In one Decision Research study, 30 percent of the participants gave odds greater than 50 to 1 that homicides are more frequent than suicides. In fact, there are about 27,000 suicides a year and 19,000 homicides. Paul Slovic is worried about this overconfidence. "It indicates that we often do not realize how little we know and how much additional information we need to know about the various problems and risks we face," he says.

Economist David Grether of the California Institute of Technology also found inconsistencies in the way people make decisions when faced with uncertainty. He has been trying to assess two phenomena experimentally: Why do people make mistakes despite the existence of established rules that make correct answers probable? And why do they then adhere so rigidly to their mistakes?

It is possible, Grether says, that each person has a set of rules-of-thumb to follow when making decisions. "We do not know what all the rules are," he says, but until more is known about how people make choices, "if I were in the business of asking people the likelihood of something happening, I would be concerned."

To find out the basic rules by which people define risks, the psychologists at Decision Research asked four different groups of people to rate 30 activities and technologies according to the risk to life they present. College students made up one group, members of the League of Women Voters a second, men and women in business and professional clubs were a third, and specialists in risk assessment were the last. Among the specialists were a geographer, an environmental-policy analyst, an economist, a lawyer, a biologist, a biochemist, and a government regulator of hazardous materials.

Decision Research found many similarities among the groups. Each believed that motorcycles, other motor vehicles, smoking, and handguns were very risky. On the opposite end of the scale, each thought that vaccinations, home appliances, and power mowers were relatively safe.

Perception gaps

There were also some strong differences. The college students and members of the League of Women Voters both rated nuclear power as the highest risk; the business and professional group rated it eighth on its list. And the group of risk assessors ranked it number 20 out of 30 items.

To determine how realistic these ratings were, the psychologists ranked the 30 items in the study according to the best available statistical evidence or estimates of annual numbers of deaths associated with them. The risk assessors' rankings were closely related to the statistical or calculated frequencies of death; it seemed reasonable to conclude that they viewed the risk of an activity or technology as synonymous with its annual fatalities. The order in which the three lay groups ranked the risks, however, showed only a moderate relationship to annual mortality rates. They apparently did not consider fatalities as a criterion or, more likely, considered them as only one among many values at risk.

Slovic argues that this does not mean the lay groups are irrational. "We call the difference between expert and lay views a perception gap. Lay perceptions are not irrational; there are reasons underlying them," he argues.

One cause of this perception gap, according to Slovic, is that specialists equate risk with the probability of death while the lay people consider many other factors.

	Members of the League of Women Voters	College Students	Members of Professional Clubs	Risk Assessors
Nuclear power	1	1	8	20
Motor vehicles	2	5	3	1
Handguns	3	2	1	4
Smoking	4	3	4	2
Motorcycles	5	6	2	6
Alcoholic beverages	6	7	5	3
General (private) aviation	7	15	11	12
Police work	8	8	7	17
Pesticides	9	4	15	8
Surgery	10	11	9	5
Fire fighting	11	10	6	18
Large construction	12	14	13	13
Hunting	13	18	10	23
Spray cans	14	13	23	26
Mountain climbing	15	22	12	29
Bicycles	16	24	14	15
Commercial aviation	17	16	18	16
Electric power	18	19	19	9
Swimming	19	30	17	10
Contraceptives	20	9	22	11
Skiing	21	25	16	30
X-rays	22	17	24	7
High school & college football	23	26	21	27
Railroads	24	23	20	19
Food preservatives	25	12	28	14
Food coloring	26	20	30	21
Power mowers	27	28	25	28
Prescription antibiotics	28	21	26	24
Home appliances	29	27	27	22
Vaccinations	30	29	29	25

Rating the risk. An ordering of risk for 30 activities and technologies, as seen by different groups, shows the perception gap that exists among them.

Paul Slovic/Decision Research

When a group of college students and members of the League of Women Voters were asked to estimate the annual number of fatalities for the list of 30 activities, each group placed nuclear power at the bottom—fewest annual fatalities—even though they had perceived it as the most risky.

How do lay people determine riskiness if not in terms of fatalities? Decision Research asked the four groups to rate the 30 items on nine characteristics that had been hypothesized as relevant to individual and societal reactions to hazards. These characteristics included the degree to which an activity's risk was voluntary, controllable, known to science, known to those exposed, familiar, dreaded, certain to be fatal, catastrophic, and imminent.

Across all 30 activities, ratings of dread and the likelihood of a mishap being fatal were closely related to lay judgments of risk; risk judgments by the League of Women Voters members and the students could be predicted almost precisely from ratings of dread and lethality. Experts' judgments were not related clearly to any of the nine risk characteristics.

In further testing, the list of technologies was expanded to include 90 items, and the list of risk characteristics was doubled, from 9 to 18. Factor analysis showed that the 18 characteristics could be represented by three factors: dread of the technology (including lack of control, lethal consequences, high catastrophic potential, and the belief that the risks associated with it are increasing and not easily reducible); how well understood the risks were perceived to be; and the number of people exposed.

Risk and regulation

When people were asked how strongly they wanted the government to control the activity, there was a strong correlation between the factor called dread, perception of risk, and desire for regulation. The higher the rated association with whatever dread represented, the higher was the perceived risk and the more people wanted the risk to be reduced through strict regulation. No relationship was found between the perceived level of understanding of the risk associated with an activity, how risky it was judged to be, or how strongly people believed it should be regulated. For example, one of the technologies whose risks were least well known was solar energy, but there was little concern about its risks or the need to regulate it. Handguns, in contrast, were the most familiar of the items, but they were also considered to be a great risk and were rated as needing regulation.

Commenting on his findings, Slovic says, "These are murky, subjective, emotional considerations. But they are important to a lot of people. The expert might say, 'Do not let your emotional views cloud the issues.' But if emotional issues are important to people, they have to be considered too."

Slovic and the Decision Research team have been in the forefront of researchers identifying public concerns that are not adequately dealt with in risk assessment. Decision makers, they believe, will have to incorporate emotional factors. However, even if the experts wanted to be sensitive to the concerns of the public, it is not clear presently how they would do it. "How do you quantify dread and integrate it with other measures of risk?" Slovic asks rhetorically. "That's a question needing further research."

Personality factors

As part of his research in understanding how new technologies—particularly those that are potentially risky—are accepted by society, psychologist Kenneth Craik at the University of California, Berkeley, is also trying to find answers to how people perceive risks. Craik, a specialist in the study of personality, is interested in knowing the interrelations among the ways a person views the risk of specific technologies, the preferred way for society to handle those risks, and selected personality variables.

"A basic premise of the field of personality is that individuals do not respond to one issue at a time," Craik explains. Instead, they usually react in an integrated manner. Craik is applying this theory to a new area—risk perception. "We are mapping it out for a better understanding of public responses to new technologies," he says.

Craik and his research team at the Institute of Personality Assessment and Research have interviewed 100 men and women in the Richmond and Oakland areas outside San Francisco. For a full day—starting with breakfast and ending in the afternoon—the participants responded to questions about 50 selected technologies, nine risk-assessment methods, and a range of personal elements—including the kinds of things that interested them, the kinds of institutions they trusted (the press? the Red Cross? the White House?), and the energy sources they preferred. The University of California team plans to repeat these interviews with people who are actively working with science and technology. This group has not been selected yet but may include industrial scientists and executives, regulatory agency staff members, science writers, environmentalists, technology assessors, and risk analysts.

Craik has not yet analyzed his data fully but finds a few intriguing suggestions. When he ran pilot interviews with Berkeley students as a way of trying out his questionnaires, he found that they wanted to know who was in charge of making a decision before they could agree on how that decision should be made. "It was difficult for them to consider a specific situation—adopting additional landing-guidance equipment in airplanes—without first attending to the question of who is responsible or who should be responsible for that sort of thing," Craik notes.

Answering the "how" question is key to Craik's research. "We have asked people to indicate the features they like about various risk-analysis methods and to rate them in terms of decisiveness, fairness, and other characteristics. We have also asked them which attributes are desirable features," he says. "In this way we will be able to identify and better understand the constituency for cost-benefit analysis or any other approach to societal decision making."

Craik also has found that the public may be changing the way it thinks about technology. When asked to rate 100 specific technologies according to how well they typify technology in general, a group of 30 participants put benign technologies such as those involved with information processing high on the list and such traditional ones as chemicals and energy further down. "We may be moving from the antitechnology days of the sixties and seventies to a more protechnology mood," says Craik. "Maybe the energy and chemical technologies added a negative element, but the new information technologies appear to be a more positive force."

The nature of risks might also be changing, according to Paul Slovic of Decision Research. Consider the rivers in Eugene, he says. They used to overflow frequently. The floods caused some damage but the community was prepared for it. Dams were built to protect against the floods. Shopping centers, homes, and running trails were built on the flood plain. There is less risk of having a flood now, but when that rare flood does occur it is likely to be very costly.

Slovic cannot answer whether things are more or less risky today. But since his research shows that risks are perceived differently by different people, it is only fitting that he responds: "It depends on which way you think life is better." •

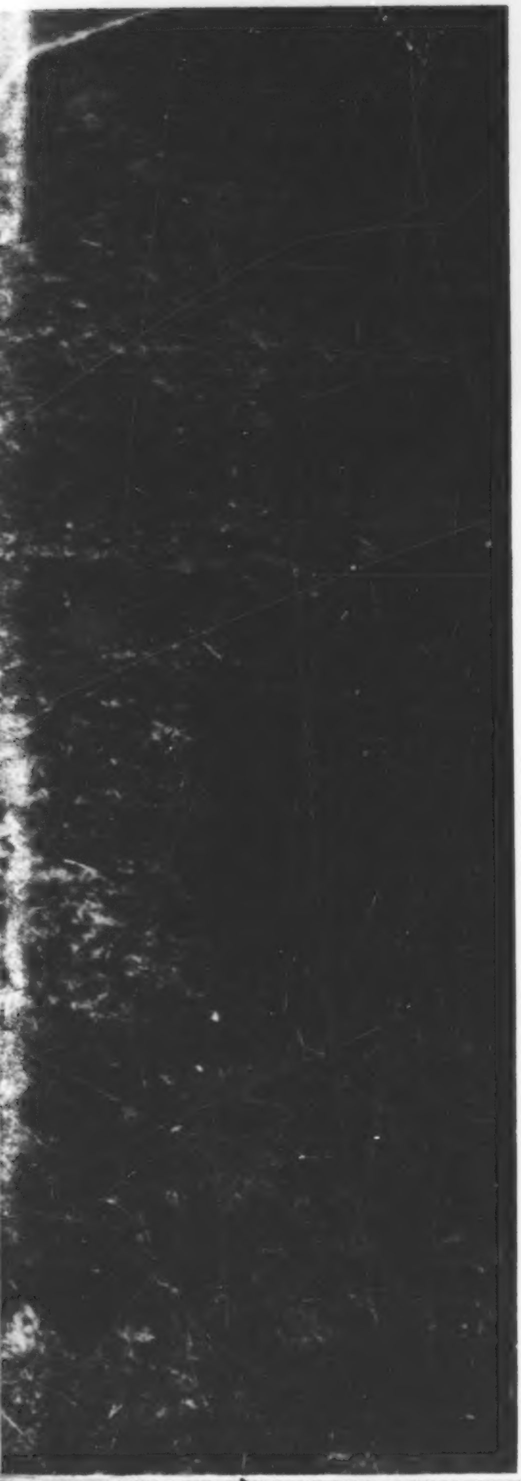
Parts of the research discussed in this article are supported by the National Science Foundation's Technology Assessment and Risk Analysis Group.

ISLANDS OF LIFE

by Sam Iker

IN A FOREST SEA

Deep in the Amazon rain forest, biogeographers are testing a theory of species survival at a time when species survival is becoming a grave problem worldwide.



Forest reserves. The first two reserves isolated in the course of the ecosystems project in Brazil. The reserve on the left is ten hectares; the plot on the right is one hectare.

R. O. Bierregaard

Word came the evening of October 29, 1980. Not long after midnight, Richard Bierregaard and his crew were heading out of the Amazon State capital of Manaus into the heart of the Brazilian jungle. Their destination—a remote *fazenda*, or cattle ranch, some 80 kilometers to the north along a bumpy dirt road.

Shortly after first light, Bierregaard and his colleagues were touching off a series of fires in swaths of dry brush and fallen trees encircling two small patches of rain forest. Three hours later, ranch workers ignited an inferno in the surrounding forest to burn off a 600-hectare clearing for new pasture land. The flames never reached the two isolated forest fragments; they were preserved by the 300-meter-wide protective corridor created by those backfires.

In the past two years Bierregaard and his fellow biologists from the World Wildlife Fund and Brazil's National Institute for Research on Amazonia have continued to carve out, or isolate, natural islands from the sea of dense tropical foliage that surrounds them. By the time they are finished, they will have fashioned as many as 60 such jungle islands, ranging from one to ten thousands hectares. And over the next two decades they plan to carry out a vast and unprecedented ecological experiment that Princeton University ecologist John Terborgh calls so massive "that it verges on the undoable." Indeed, says Terborgh, "it's exciting to think that anyone would even try anything so grandiose."

In the coming years the researchers will be observing the changing ecology of these terrestrial islands, monitoring closely what happens to the birds, animals, insects, and plants. The results of their studies could be vital in helping to manage and preserve the unique biological resources of the vast Amazon rain forest. At the same time, what they learn promises to clarify some of the questions surrounding a controversial ecological concept called island biogeography.

Shrinking habitat

Elements of the intriguing theory are reflected in the sweeping U.S.-Brazilian experiment. They are also central in Brazil's efforts to establish an extraordinary system of biological reserves and parks in the Amazon region, by far the largest expanse of tropical rain forest left in the world today.

Scientists warn that these tropical forests, which contain an estimated two-thirds of the planet's 4.5 million animal and plant species, are disappearing with alarming speed. They calculate that such forests are being cleared at the rate of nearly 11 million hectares yearly, an area about half the size of California. And fears are that the pace will, if anything, increase under the pressure of a food- and energy-short world population whose numbers will double by the year 2000.

The implications are sobering. Says Ira Rubinoff, director of the Smithsonian Tropical Research Institute, "If deforestation continues at the present rate, we may expect to witness the disappearance of almost a million species—many of which have potential benefits to humans as sources of domesticated plants and animals, medicines, fibers, and oils—which will be lost forever with the forests."

This was the situation facing scientists and planners back in the mid-1970s when they confronted the task of preserving the Amazon's biological richness. The question, recalls Gary Wetterberg, one of the planners who is now with the international unit of the National Park Service, was "how, from a biological perspective, do you make sense of what appears to be a homogenous mass of forest on most world vegetation maps?"

The planners decided to focus on so-called Pleistocene refugia, areas where concentrations of endemic species were clustered. It was believed that such endemic species evolved during the Ice Age, when the Amazon basin was cooler and drier and the region

Saving some. Backfires protect designated reserves from being consumed when the surrounding area is burned out to prepare it to become pasture.

R. O. Bierregaard

was not completely forested. Scientists had identified refugia for certain species of birds, lizards, woody plants, and butterflies. These locations were mapped on separate overlays. "Where two or more areas of endemism coincided," says Wetterberg, "was a priority area."

Optimum sizes

The question then became one of size. What is the minimum space needed to preserve and sustain various species? For some, such as plants or insects, obviously little area is necessary. But an eagle or other large bird may require 10 to 20 square kilometers to survive, and a large predatory mammal such as a jaguar needs perhaps 30 square kilometers of habitat to sustain itself.

Still, little systematic reseach had been published on this key issue. One of the few who had focused on the problem was Princeton's John Terborgh. Based on studies involving principles of island biogeography, he concluded that neotropical bird species in lowland rain forests required "on the order" of 2,500 square kilometers to hold extinction rates to "less than 1 percent of the initial species complement per century."

According to Gary Wetterberg, Terborgh's work gave "by far the largest figure published up to that time for minimum necessary size with good reasoning behind it. So," adds Wetterberg, "we basically took his recommendation and doubled it in size, adding a buffer zone as a safety factor." The planners then set about mapping 24 "major conservation units"—each containing a core area of some 5,000 square kilometers or more. Since then, more than 20 Amazonian reserves and parks have been officially established, most of them well over the minimum of 5,000 square kilometers called for in the initial blueprint.

Terborgh himself is quick to note that his 1975 "minimum-area" calculation applied only to neotropical birds. "I'm sure the result would be very different for other groups of organisms," he says, "but we just don't have the data at this point to draw the inferences." The model from which Terborgh gauged his estimate for neotropical birds involved islands in the Carribean. "I'd love to see other people follow it up," he adds, "with similar analyses for other groups of organisms or other islands."

Such knowledge could help conservation planners slow the rising tide of extinctions. As the world's wilderness areas rapidly shrink, more and more of the earth's wild flora and fauna will be confined to parks and refuges surrounded by inhospitable habitat. In a sense, these reserves will be like terrestrial islands in a hostile sea. And the question is: What happens to such isolated natural communities?

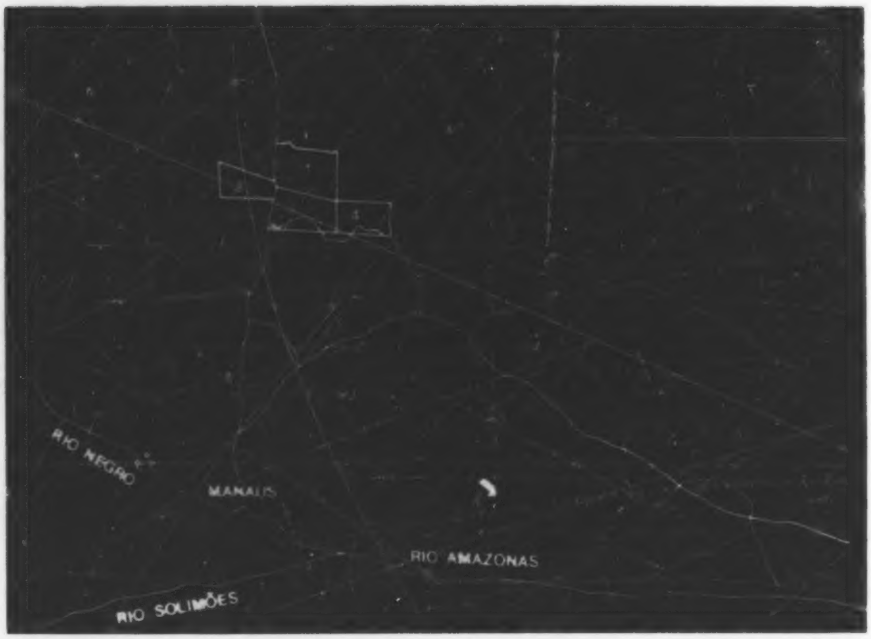
Some 20 years ago, mathematician Robert MacArthur of Princeton and Harvard entomologist Edward Wilson evolved one answer with their equilibrium theory of island biogeography. In their theoretical model the number of species on an island is determined by two factors—extinctions and in-migration. Extinctions are influenced by the area of an island: the smaller the area, the higher the extinction rate. The in-migration of new species varies with the distance of an island from the mainland: the closer the island, the higher the in-migration rate.

Ultimately, according to MacArthur and Wilson, the number of species on an island will reach a stage of dynamic equilibrium, with species extinctions being roughly balanced by the in-migration of the same or other species suited to available ecological niches. This equilibrium level will be highest on large islands that lie closest to the mainland and lowest on small islands that are far out in the ocean.

The MacArthur-Wilson hypothesis, which contained long-known elements such as the relationship between species numbers and area size, aroused intense interest among ecologists. Notre Dame University's Robert

---

lker writes frequently on environmental and ecological subjects.



Project site. Four adjoining farms in the agriculture/ranching district north of Manaus in Brazil are the site of the minimum size ecosystem-preservation project.

World Wildlife Fund

McIntosh, an authority on the history of ecology as a science, explains that island biogeography "was one of a number of theoretical propositions which came on during the late 1950s and the 1960s. It was widely held as the new wave in ecology: presumably an application of real theoretical, deductive science, which hadn't previously been well marked in ecology."

Seeking data

As researchers set about testing the new model, the definition of island was further refined to include any isolated biological community. Aside from focusing on real islands in the accepted sense, scientists studied mountaintops, coral heads, fragments of forest land, and even, in some regards, continents. Some created artificial islands, including microscope slides of various sizes inhabited by aquatic algae. Other biologists even speculated that individual animals could be considered islands; one way proposed to investigate this thesis was to study the parasites (as many as nine species of worms) that live in the intestines of bobwhite quail.

Some of the most persuasive evidence pointing to a dynamic equilibrium phenomenon grew out of a carefully controlled experiment conducted by Wilson and a youthful biologist, Daniel Simberloff, in 1968. The researchers focused on a cluster of six tiny mangrove islands off the Florida Keys. First they did a census of insects inhabiting the islands. Then, in a complex logistical operation, they exterminated the insects.

During the following months, the scientists regularly monitored the recolonization of the mangrove islets. After about eight months, they found that virtually all six had been reoccupied. The number and composition of the returned insect species were similar to those of nearby, untreated mangrove islands. The result, according to Simberloff, suggested "a true equilibrium."

Many ecologists took another approach. They analyzed species numbers and extinction trends on actual islands that seemed to be in equilibrium, then studied others where the species numbers appeared to be declining toward such a balance. Most of the research involved birds since they were the easiest creatures to observe and count. Moreover, in some cases, historical bird counts provided baseline data.

Islands in equilibrium are in general oceanic—located far from any land mass. Those in a transition stage are so-called land-bridge islands. These would have been cut off from nearby continents at the end of the last large glaciation, some 10,000 years ago, as the ice melted and sea levels increased. When they were separated from the mainland, they contained more species than their reduced areas could accommodate. Over thousands of years, according to the theorists, the number of species on such saturated land-bridge islands has been declining toward what scientists consider a rough state of equilibrium.

In testing the concepts of island biogeography, ecologists compared the number of bird species on oceanic islands with the number of land-bridge islands. They found, for instance, that large land-bridge islands were populated by far more species than equivalent-sized oceanic islands. (The distinction did not seem to hold for small islands.) Studies of archipelago islands in the Southwest Pacific showed that those farthest from the mainland or from nearby large islands had fewer non-marine bird species than those close to a major land mass. The observations seemed to confirm the theory that in-migration drops as the distance from a pool of species grows.

Toward a model

Extinction rates, according to the MacArthur-Wilson concept, are closely related to area. The loss of species will be much more rapid on small islands than on large ones. The hypothesis is that populations of individual species on large islands are higher. Thus various species are less prone to extinction from inbreeding, competition, disease, natural disasters, and other causes.

Given these observations (supported by considerable research), the next step was to develop a predictive model that would forecast extinction curves. Princeton's John Terborgh started with the broad assumption that land-bridge islands originally contained generally the same number and assortment of species as comparable areas on the mainland. Using five Caribbean land-bridge islands ranging from Isla del Rey, at 249 square kilometers, to Trinidad's 4,828 square kilometers, he then considered several theoretical models.

The one that seemed most logical held that the extinction rate was proportional to the square of the number of species originally present. That would mean, explains Terborgh, "that the extinction rate will be rapid at the start, when there are a lot of species, and will slow down as species drop out of the system. So the survival potential of the remaining species improves over time. This model, on the basis of *a priori* reasoning, was the most appealing and biologically realistic. It also fit various existing data sets better than the other two models."

To test his model, Terborgh looked at the little island of Barro Colorado, which has been isolated in Lake Gatun when the Panama Canal was built. The island was declared a biological reserve, and its wildlife

Building forest islands. Thomas Lovejoy (left) and Richard Bierregaard (shown freeing a motmot from a mist net in an inventory exercise). Their experiment—the creation of island habitats in cleared Brazilian jungle areas—is to see if such devices can be used in species preservation.

World Wildlife Fund; T. E. Lovejoy

has been intensively studied. As a result, there are detailed records of extinctions that have occurred since 1923.

Much of the species loss since then resulted from habitat changes that took place after the island was cut off. But Terborgh concentrated on the 15 species that had disappeared from unaltered forest areas. Given the 160-hectare area of the island, the initial census count of 205 species, and the elapsed time of 50 years, his model predicted the extinction of 16 or 17 species. The result, declares Terborgh, "is strikingly in agreement with what has been observed on Barro Colorado."

Terborgh believes that the "analogy to future parks cut off from dispersal, and land-bridge islands, is an appropriate one." But he cautions that the two are not equivalent. "Many species which live in the kinds of habitat that isolate remote woodlots," he notes, "are able to invade the lots. So in-migration rates of alien species are very high, whereas in a true island they're very low. This higher in-migration threat to a habitat island could increase the rate of extinction above that which occurs on an island isolated by water."

Nevertheless, there are indications that trends on land-bridge islands are similar to those operating in isolated patches of continental habitat. The strongest evidence, according to Terborgh, comes from ornithologist Edwin Willis's study of three forest remnants in the Brazilian state of São Paulo. "The area is otherwise 99.9 percent deforested," says Terborgh. "The whole landscape has been transformed by human activities."

Based on the gap between early bird counts and recent observations, Willis found that a 1,400-hectare patch of forest had lost only 14 percent of its breeding bird species. The figure for a 250-hectare fragment was 41 percent. And in the smallest patch, 21 hectares, 62 percent of the resident bird species had disappeared. From these and similar findings, Terborgh concludes, "we must have reserves in the size range of from one to several thousand kilometers if we're to prevent a veritable rush of extinctions."

Questions of merit

This conviction is shared by like-minded colleagues such as Willis, Jared Diamond of the University of California at Los Angeles, and others. But it is not universally shared. There are scientists who argue that a series of smaller refuges may be as effective or perhaps better for maintaining maximum species diversity. Moreover, they challenge not only the usefulness of island biogeography as a theory but also the validity of many follow-up research studies.

The most prominent and outspoken critic, ironically, is Florida State University biologist Daniel Simberloff, who with Edward Wilson conducted the best-known controlled island-biogeography experiments in the late 1960s. Simberloff agrees that his early work with Wilson indeed suggests that equilibrium and a species-area relationship does exist. But he questions later "inferential" research that fails to test the theory directly.

Simberloff disputes some of the assumptions that underlie many island biogeography studies. He voices doubt, for instance, that it is possible to estimate how many species occupied a land-bridge island ten thousand years ago. "It's assumed," he says, "that this one little patch of land, which is now an island, originally contained all of the species of the contiguous land mass. But in any quadrant on the mainland you may find a complete set of species or only a few species. Also, there's been substantial habitat change in the interim which we know very little about. That in itself would affect the species composition and richness of any one of these sites. So there's no way you can really say anything about what was on an island in the beginning or what trajectory through time the number of species on that particular patch took."

What about the current bird counts which seem to show that some species have gone extinct? Again, Simberloff and others have their doubts. They contend that censuses were taken at different times by different people, and they vary in quality. Moreover, many extinctions are not really extinctions at all but rather transient movements of bird populations between islands or between island and mainland. In other words, they maintain, the absence of a particular species at the time a bird count was taken does not necessarily mean that the species is extinct on an island.

"It's not that people set out to do bad work," declares Simberloff. "It's just that a fair test of the theory requires data that are almost impossible to get." As far as Simberloff is concerned, "island-biogeography theory has very little to offer. It's all things to all people. By subsuming all results, it loses its predictive or explanatory power."

Not surprisingly, John Terborgh dismisses such criticism. He says that while Simberloff and others attack the "methods and details of procedures" that he and his fellow island biogeographers use, they "seldom, if ever, show that our conclusions are wrong. There's an important distinction to be made."

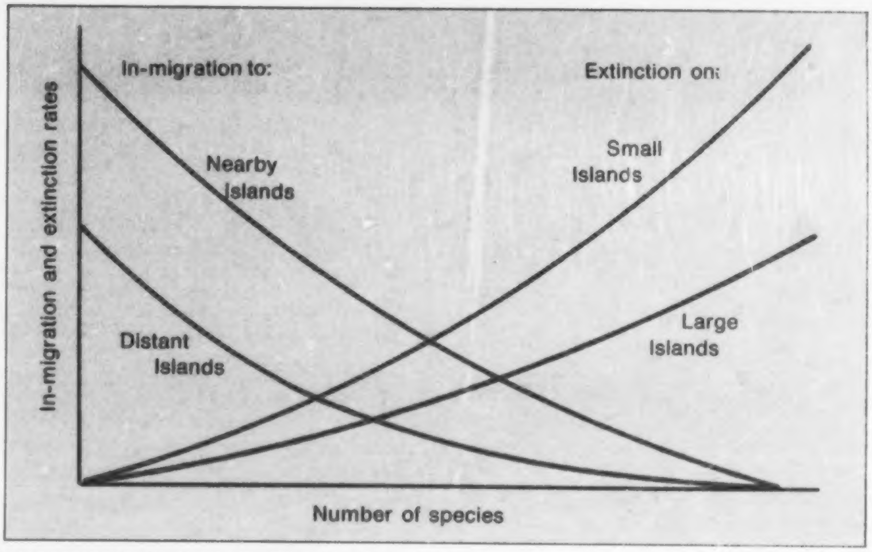
According to Terborgh, the critics are a minority who "don't like to acknowledge that natural experiments are in any way valid or replicable. But there's abundant evidence to the contrary—that nature itself has performed replicate experiments and the results are often quite consistent. There are numerous archipelagos where nature has provided the replicates in the form of closely adjacent and very, very similar islands."

The heated debate, according to ecology historian Robert McIntosh, involves "substantial issues." And, he adds, "some very prominent ecologists staked their reputations on it, one way or another." What's going on, he explains, "is a period of strife that often occurs in science. Here's a scientific concept in the middle of a controversy between perfectly competent authorities on each side. At the moment, I can't point to any synthesis that would provide a definitive answer."

Answers from the Amazon

Some of the answers, however, could be provided by the massive ecological experiment being conducted in the Amazon rain forest by the World Wildlife Fund and the Brazilian Institute for Research on Amazonia. Creating dozens of habitat islands of varying size and monitoring them over decades should finally settle one controversy. According to National Wildlife Federation scientist Thomas Lovejoy, architect and co-director of the project, "We'll know, for certain groups of organisms at least, whether in forest environments you end up with more species in a series of small reserves or in one big reserve of the same total area."

The researchers will also learn much more about the phenomenon of turnover, the change in composition of species as an island's organisms decline toward a stage of equilibrium. Such a change "would be the first sound test of turnover rates, in my opinion," says Simberloff.



Biogeographic theory. In the theory as described by Robert MacArthur and E. O. Wilson in 1967, equilibrium among species on an island is determined by island size and island distance from a species pool. Equilibria occur where the extinction and in-migration curves cross.

After M. Williamson, *Island Populations.*

The biologists should gain new insights into the mechanics of extinction. As Terborgh puts it, "We don't know very much about why species go extinct on isolated islands or in forest remnants. All the mathematics does is describe the rate at which they go extinct." Scientists do know that species with the smallest initial populations on an island are the most vulnerable to extinction, along with those whose habitat and food requirements are the most specialized. Large predators in need of wide foraging areas, for example, have low population densities and are more prone to extinction.

Over time, the Amazon experiment could reveal the patterns of extinctions in greater detail than ever before. Are species lost in a random pattern in terms of numbers and sequence? Is there a predictable sequence of extinctions, producing similar species composition in roughly equal-sized forest islands? Or is there a predictable sequence, resulting in differences in the species makeup of different forest patches, depending on which species were originally present?

These are the kinds of questions that Lovejoy and his collaborators hope to answer. Unlike the island biogeographer, however, Lovejoy's main conservation focus is to determine the minimum critical size of an ecosystem rather than the size of a refuge needed to preserve maximum species diversity. In the real world, the two may be the same.

"There are two ways of looking at the problem," Lovejoy explains. "One is asking how you design a national park so it'll hold the greatest possible number of species. What I'm talking about is rather the elusive goal of defining the minimum size needed to maintain the characteristic diversity of an ecosystem over time. In other words, I think the goals of conservation aren't simply to protect the full array of plant and animal species on the planet, but rather also to protect them in their natural associations so that the relationships between species are preserved and the evolutionary and ecological processes are protected."

The relevance of the Amazon project to island biogeography theory is far from casual. "As soon as you fragment one continuous habitat," declares Lovejoy, "the isolated fragments begin to shed species and decline to a simpler ecosystem. Two things are related: that decay function and whatever drives it. They are what define the minimum area needed to maintain the characteristic diversity. The notion of this project is to look at that decay function in an experimental way so you can follow it from beginning to end.

"If we can gain some notion about what drives the decay function," Lovejoy adds, "if there's any order to which species are lost so that a piece of forest of a particular size always tends to end up with the same remnant collection of species, then we've gained some ability to manage an isolated patch so it will hold things it otherwise might not."

To Lovejoy, the Brazilian plan for parks and refuges drawn up by Gary Wetterberg and others back in the mid-1970s presented a unique opportunity. In much of the island biogeography research, he notes, "more often than not the investigator was stuck with a situation of knowing what was on an island at the present but not knowing what was there at the moment of isolation. It was necessary to deduce how it got to where it was from a hypothetical starting point. [The Amazon rain forest] was a system where you actually could go at the thing from the very beginning and eliminate some of that fuzziness. It was immensely attractive."

A fitting arrangement

In what Lovejoy calls "an extraordinary coincidence," everything fell into place: the development of the new Brazilian parks plan; a willingness of Brazilian authorities to join in the research effort; the existence of a modern scientific institute in Manaus; a "virtually untouched rain forest with jaguars, harpy eagles, pumas, tapirs, macaws, and all the rest" just two hours away from Manaus via a new road to Venezuela. Added to these are a Brazilian law requiring that 50 percent of any development project in the Amazon be left in forest, and the willingness of cattle ranchers in the area to cooperate by leaving specific patches of jungle untouched as land was cleared for pasture.

Starting in 1980, the American and Brazilian scientists began carving out their landlocked islands from the virgin rain forest—a series of 1-, 10-, 100-, and 1,000-hectare plots. After surveying and staking out a roughly square patch of forest, the researchers carefully inventory the target flora and fauna, including birds, primates, other mammals, insects, and trees. Ranch crews then cut down the understory foliage and trees surrounding the future island. The felled brush and trees are allowed to dry. "When they judge the time is ripe," says Lovejoy, "they set a match to it and off it goes." To protect the small-hectare plots from the inferno, the researchers set backfires to create insulating corridors.

"The reason for having a series of different-sized reserves," explains Lovejoy, "is to gain a more comprehensive insight into what's going on. The one-hectare reserves are going to go through changes very rapidly. Their decay function is probably a matter of several months. The ten-hectare patches are going to change more slowly and in a more complicated way. And the same is true for the hundred- and thousand-hectare reserves.

Various sized plots will be clustered and, as far as possible, replicates of each other. That will allow the scientists to see how predictable changes in species composition are in terms of area. As a control for comparison purposes, the researchers will have both their baseline inventory data and a 10,000-hectare reserve where the decay processes occur at a much slower rate.

The project investigators—permanent staff, visiting scientists, and graduate students—alternate between Manaus and camps at the jungle site. Field work is under way almost daily. Ornithologists are doing bird censuses in each plot at three-week intervals. Other scientists are collecting data on ants, butterflies, reptiles and amphibians, primates and other animals, and woody plants.

Some 20 habitat islands have been created since 1980. And, says Lovejoy, "we've already learned some interesting things." One phenomenon documented for the first time is termed influx. When the forest surrounding the first two plots was felled by ranch workers, the bird-capture rates suddenly doubled as wildlife from the destroyed habitat fled into the intact refuges.

In the one-hectare reserve, this elevated recapture rate fell below the preisolation level within 12 months. And as the biologists expected, the capture curves are gradually falling as the ecosystem decay process proceeds. The researchers noted other changes.

Within two to four months after isolation, two species of birds that follow and feed on swarming army ants had vanished. Both species need a considerable area to search out their army-ant prey. Other species, including fruit-eating birds, which feed below the forest canopy, had also largely disappeared, along with a band of golden-handed tamarin monkeys. Two saki monkeys, which feed on fruit and range over hundreds of hectares, vainly searched for fruit-bearing trees and then vanished. Howler monkeys, which eat foliage, seem to be adapting to the isolated ten-hectare reserve.

Such early results, says Lovejoy, "show how complex the dynamics of forest fragments can be. One of the advantages of the project is that it's looking at the problem in perhaps the most complicated biological situation possible. In one way, it makes the job more difficult. In another sense, it may reveal certain aspects of species-loss biology that wouldn't be so apparent in a simpler system." •

The National Science Foundation contributes to its share of the research discussed in this article through its Ecology Program.

Superconductivity and magnetism: an uneasy alliance

by Peter Gwynne

The incompatibility of superconductivity and magnetism once was a given. Under certain conditions and in certain materials, marriages can be arranged.

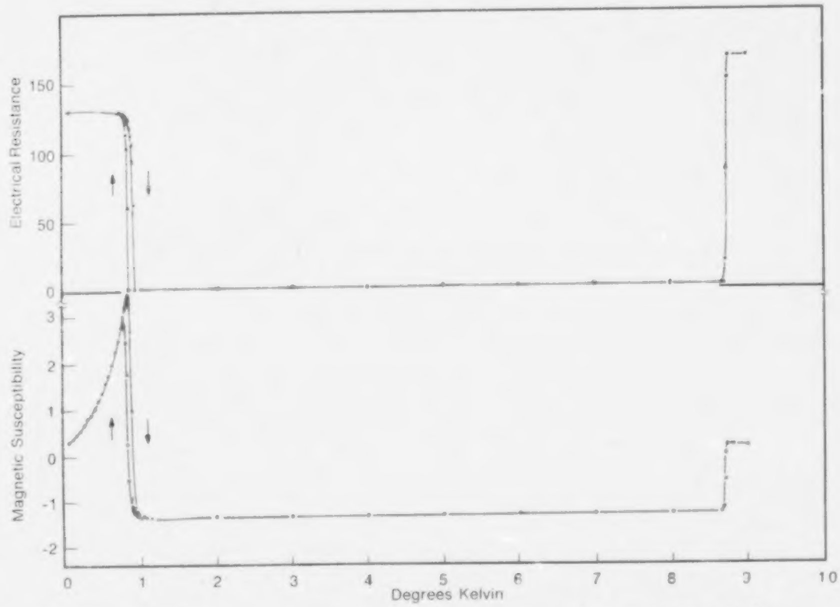
hysicists used to regard superconductivity and magnetism in the same way that they looked on oil and water. The two just wouldn't mix. This apparently need not be the case. A combination of complex experimental observations and bold theoretical advances has convinced many scientists that at least a few unusual alloys can exist simultaneously in the states of zero electrical resistance and measurable magnetic intensity. The marriages that are being achieved, however, are uneasy ones at best.

The uncomfortable alliance between two naturally opposed characteristics of matter may appear arcane. Yet the effort to determine on an atomic level just how the competition between superconductivity and magnetism resolves itself is providing fresh and vital insights into the fundamental nature of solids. Entirely new phases of matter, internal contests among different varieties of magnetism, and new knowledge of the critical phenomena that are always present when one form of matter changes into another represent just a few of the frontiers on which this new field of basic research is pressing.

Theoretically, the subject touches on many of the underlying concepts of solid-state physics. At the base of both superconductivity and magnetism is the idea of order; in the scientific sense, the term applies to regularity more rigorously than it does in normal usage.

Order in matter

Imagine an ordinary crystal, whose individual atoms take up fixed positions in what crystallographers term the lattice. If one starts at a specific atom and moves through the lattice, one will encounter atoms (or ions and molecules) at regular intervals. The spacing between adjacent atoms in the crystal does not change according to their position in the crystal. Hence, physicists say, the crystal possesses long-range order.



Temperature response. Superconductivity and susceptibility to ferromagnetism in the compound erbium rhodium boride appear and disappear together in response to temperature change. Both quantities decline abruptly when the temperature drops below 8.7 degrees Kelvin; they increase just as rapidly at a lower transition temperature of 0.9 degree.

M. Brian Maple

A liquid is different. Its atoms jumble together in an irregular array. Setting out in a straight line from any specific atom, one might travel long or short distances before encountering any other atom. There are, in general, no consistent repeat distances among atoms. Plainly, liquids have no long-range order. However, when liquids are only a few degrees above their freezing points, their atoms tend to clump together in small assemblages that show a certain amount of regularity. Within a volume containing just a few tens or hundreds of atoms, it is possible to detect repeat distance. Such assemblies as these illustrate what physicists call short-range order.

The idea of order does not apply only to the positions of atoms. One vital property that can possess order or disorder is spin. Mathematically, spin is an extremely complicated concept. However, it can be regarded simply as a manifestation of an electron's revolution around its own axis. According to quantum mechanics, an individual electron can spin in just one of two ways—pictured as a clockwise and an anticlockwise revolution, as viewed from above. Electron spins in a solid are said to be ordered if they correlate over a long range—if all are clockwise, for example, or if alternate spins are clockwise and anticlockwise.

Order and magnetism

Another property of solids to which physicists apply the idea of order is the magnetic moment. This characteristic indicates the magnetic strength of any solid, atom, or electron. The size of an electronic magnetic moment depends on the electron's spin and on its orbital motion around the atomic nucleus to which it is attached. The magnitude of the magnetic moment of an atom stems largely from the spins of the atom's outermost electrons. If the spins of all the outside electrons are aligned, the atom has a larger magnetic moment that if would if spins pointed in opposite directions.

All elements and compounds possess some form of magnetism. Most types are not permanent, however. Substances that are paramagnetic and diamagnetic, for example, can never be induced into a state in which they magnetically attract other substances. Since most substances possess one of these two types of magnetism, manifestations of what most people recognize as magnetic effects are seen only rarely in daily life.

Recognizable magnetism is the permanent form, known technically as ferromagnetism. It is responsible for the horseshoe magnet's attraction for iron filings and the large-scale generation of electricity. Only a few metals possess ferromagnetism at normal temperatures—most notably iron, cobalt, and nickel, and some of their alloys, and rare-earth metals.

Ferromagnetism stems from order among the metals' magnetic moments. The magnetic moments in relatively large segments, or domains, of ferromagnetic substances all point in the same direction. This ordered behavior can be destroyed by heating the metal above the temperature characteristic of the metal or alloy and known as its Curie point.

Order and superconductivity

Solid-state physicists recognize another important type of permanent magnetism, which they term antiferromagnetism. The magnetic moments in antiferromagnets point alternately in opposite directions. Thus, even over the scale of a few atomic separations, the moments cancel themselves out. Because of this, a crystal of an antiferromagnet has no ability to attract objects magnetically. Nevertheless, the crystal is completely ordered over the long range in the magnetic sense. Antiferromagnetism can be destroyed in the same way as ferromagnetism; the characteristic temperature is called the Néel point.

Superconductivity also stems from order inside metallic elements and alloys. In metals, the outer electrons of atoms do not stay rigidly close to their atoms. Instead, they merge together in a common pool, which can drift through the metal. By moving in any particular direction, the free electrons carry electric current. In most metals and alloys, the flow of free electrons is hindered by collisions caused by vibrations and defects in the ordered array of ions—the atoms that have lost electrons. The collisions produce the metals' resistance to the flow of electricity. A few metals and alloys at very low temperatures, however, show no resistance to the passage of electrons. These are superconductors, and order is also involved.

Discovered in 1911 by H. K. Onnes, superconductivity gained its theoretical underpinning with publication in 1957 of the BCS theory, named for John Bardeen, Leon Cooper, and John Robert Schrieffer. The theory envisions the free electrons in superconductors as paired so the electron cloud can sweep through the crystal lattice of ions unimpeded. One particular display of ordering in superconductors is the existence of so-called Cooper pairs of electrons. They are pairs that have linked together because they have exactly equal momentums and spins—pointed in opposite directions.

Solid-state physicists have identified two different varieties of superconductor, called type-one and type-two. In the first, the superconducting ability disappears when the metal is placed in a magnetic field of relatively low strength—just a few hundred

---

Gwynne is former science editor of Newsweek.

Matchmakers. The late Bernd T. Matthias and M. Brian Maple worked together in 1977 (top left). Maple and Steve Lambert, loading a sample into cryostat (top right), and Jeffrey Lynn (below) are studying the characteristics of different compounds.

A. C. Lawson; Al Danegger

gauss, equivalent to that in a small bar magnet. Superconducting elements such as tin and lead are this kind. Type-two superconductors, by contrast, can withstand very high magnetic fields. Experimentalists have applied fields as high as several hundred thousand gauss to a few superconductors before their electrical resistance started to appear. But there are thresholds.

Why do magnetic fields eventually destroy superconductivity? Theorists working a quarter of a century ago concluded that somehow the fields broke up the Cooper pairs of electrons. But was this a necessary consequence whenever magnetism and superconductivity were mixed, or could at least some magnetic compounds exist in a superconducting state? The same theorists decided that coexistence of the two properties might occasionally occur, and they set out to make specific predictions.

Soviet physicists L. P. Gor'kov and A. I. Rusinov, for example, suggested that ferromagnetism and superconductivity might be compatible and predicted a phenomenon called reentrant superconductivity: As its temperature fell, a compound or alloy, capable of reentrance whatever its constituents, would first become superconducting and then lose its superconductivity in the vicinity of the Curie point at which it became ferromagnetic. It remained, however, for reentrant compounds to be found in nature. Gor'kov and Rusinov also predicted that under certain conditions, superconductivity would persist below the Curie point in a ferromagnetic superconductor.

Phillip Anderson of Bell Labs (to become a Nobel prizewinner in physics) and Harry Suhl, now at the University of California at San Diego, predicted the existence of small zones of compatibility between ferromagnetism and superconductivity. Since the predicted zones would be only a few angstroms in size, Anderson and Suhl pointed out that they would not be examples of truly ordered ferromagnetism. They coined the term "cryptoferromagnetism" for their prediction.

Twilight zones

The first group to seek experimental evidence of the theories was headed by Bernd Matthias of Bell Labs, one of the legendary figures in experimental physics. Over the years, Matthias had discovered dozens of new superconducting compounds, with increasingly higher transition temperatures marking the disappearance of their superconductivity. (See "Superconductivity's Material Question," *Mosaic*, Volume 10, Number 1.)

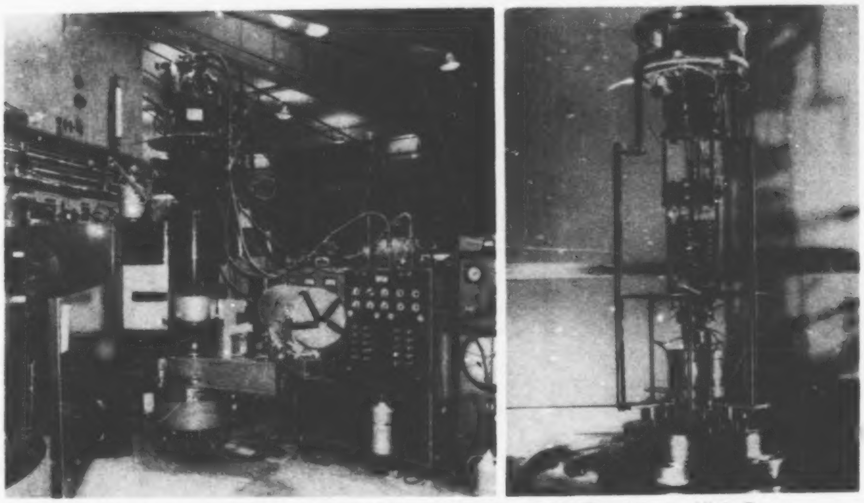
Starting the search for compatibility between magnetism and superconductivity, Matthias dissolved small amounts of rare earths—the series of obscure elements that runs from lanthanum to lutetium in the periodic table—in various superconductors. Most of the rare earths are magnetic. Dissolving rare-earth ions into superconductors introduces internal magnetic fields in them.

Generally, the effect of the ions, Matthias found, was to reduce spectacularly the transition temperatures above which the superconductors lost their resistance-free character. However, a few alloys showed only a modest decline of the superconducting transition temperature, and the magnetic ordering temperature measured when percentages of rare-earth ions increased suggested an interesting possibility: There might be "twilight zones"—regions, Matthias reasoned, in which magnetism and superconductivity could coexist.

Checking the hunch was difficult for two reasons. The percentages of rare earths at the twilight zones were small, which meant that the ions would be distributed throughout the alloys in random rather than ordered fashion. Further, because of metallurgical incompatibility in at least some of the mixtures, it seemed likely that the rare-earth ions were clumping. The consequent clusters could create islands of magnetism in a sea of superconductivity. Because analytical techniques were not very refined at the time, "it just wasn't possible to draw definite conclusions and say whether magnetism and superconductivity really coexisted," recalls M. Brian Maple of the University of California at San Diego, who had worked with Matthias. "But in retrospect there were no real magnetic-superconducting combinations."

A few years later, in the mid 1960s, experimentalists wandered up another blind alley. It involved the Kondo effect, named for a Japanese physicist who identified an interaction of the spins of superconducting electrons with impurity ions added to a superconductor. As the electrons swing past the ions, their spins tend to line up either parallel or opposite to those of the ions. This tendency would oppose the electron spins' natural inclination to retain their orientations and would inhibit the Cooper pairs that are responsible for the superconductivity.

In some alloys (varied percentages of the metal cerium added to an alloy of lanthanum and aluminum, for example), the competition leads to reentrant superconductivity. As its temperature falls, the alloy first becomes superconducting and then returns to a normal state of electrical resistance. Interesting as the effect is, though, it happens at a strictly local level—at zones in the material. "It has nothing to do with long-range magnetic order," emphasizes Maple.



Equipment. The experimental neutron-scattering equipment at the Brookhaven High-Beam Reactor (left) and the inside of a helium dilution refrigerator used to achieve ultra-low temperatures at the University of Maryland (right).

Brookhaven National Laboratory; Jeffrey Lynn/University of Maryland

The quest for true compatibility between magnetism and superconductivity took a fresh turn in the early 1970s. The stimulus was the discovery by a French metallurgist of what are now called Chevrel phases. These take the form of one of a range of metals alloyed with molybdenum, sulfur, and selenium.

Order reordered

In 1972, Bernd Matthias discovered that certain Chevrel systems were superconducting. Shortly after, Oystein Fischer, a Norwegian physicist at the University of Geneva, determined that the superconducting Chevrel systems could retain their zero resistance in extremely high magnetic fields—as much as 500,000 gauss. Then in 1975, Fischer and others reported that they had made Chevrel phases with rare earths—and that such compounds were superconducting.

Because most rare earths possess permanent magnetism of one type or another, the scene was set for another look at the possibility of compatibility. What experimentalists needed were accurate means to prove the existence of long-range order in the compounds, and reliable methods of producing the rare-earth Chevrel compounds.

The major difficulty in making the compounds comes in hitting the exact chemical composition. The problem of islands of magnetically ordered material in a superconducting sea, which tantalized the early experimenters studying compatibility, does not arise in the Chevrel phases; all are physically ordered, with specific molecules or atoms occupying specific sites in the crystals. However, because the Chevrel systems all involve three elements, the possibility of slight impurities that can skew the magnetic measurements is ever present.

"So far, I have never seen a Chevrel-phase compound without an impurity phase," remarks Brian Maple. Rare-earth oxides, sulfides and oxysulfides, molybdenum sulfides, free molybdenum, and rare earths are just a few of the impurities that somehow worm their way into the most meticulously prepared Chevrel compounds.

Careful handling can minimize the effects of impurities. There are several techniques: grinding the starting materials into a fine powder prior to heat treatment to reduce the chance of including filaments of superconducting impurity phases; measuring properties (such as specific heat) that are sensitive to the presence of bulk superconductivity; and powdering the final products to minimize the contribution to electrical and magnetic measurements of any remaining impurities.

Once the Chevrel compound is satisfactorily manufactured, its magnetic order must be assessed. The occurrence of antiferromagnetic order in the superconducting state in the Chevrel-phase materials was first established by means of bulk measurements such as electrical resistivity, magnetic susceptibility, specific heat, and features in plots of the upper critical magnetic field as a function of temperature. However, it is difficult to determine the detailed nature of the magnetism of these substances by orthodox magnetic measurement techniques because of the superconductors' tendency to expel the applied magnetic fields in which the measurements are made. Moreover, it is also desirable to measure moments in extremely localized areas, around individual atoms.

One promising approach is the Mössbauer effect. Discovered in 1962 by a German scientist, the effect is the recoilless emission by certain atomic nuclei of powerfully penetrating, high-frequency gamma rays. The wavelengths of the emitted gamma rays depend in part on the local magnetic fields around the nuclei that emit them. Hence, the technique can in theory pinpoint magnetic order within very small regions.

Critics assert that Mössbauer measurements are often difficult to interpret. However, Gopal Shenoy of the Argonne National Laboratory, a prime exponent of the technique, says that the problems have been ironed out. "We have tried hard to educate the others," he says. "As far as interpretation of magnetism is concerned, we have no problem with the Mössbauer effect."

Most experimenters, however, use more costly but more direct means to study magnetic order in superconductors. In a perfect world, they would prefer to use X-rays. Unfortunately, X-rays have no magnetic moment; they hardly react at all in a magnetic fashion with atoms. On the other hand, neutrons, the neutrally charged particles that share the atomic nucleus with protons, do have magnetic moments. Thus when neutrons are fired at crystals, they produce two sets of images. One provides clues to the crystals' spatial structure, as do X-ray images. The other gives data on the crystals' magnetic makeup.

Neutron scattering

"In typical magnetic experiments like measuring magnetic susceptibility or magnetic moment, you cannot tell what is going on inside the crystal," says Jeffrey Lynn of the University of Maryland. "Thus neutron scattering really gives you unique information." Lawrence Passell of the Brookhaven National Laboratory puts it more succinctly. "Neutron scattering," he enthuses, "is as perfect a probe as you can imagine."

Brookhaven's High-Flux Beam Reactor is one of a handful in the United States dedicated to neutron diffraction studies. The reactor produces streams of neutrons of different energies that emerge from eight portholes grouped around the reactor. To prepare a beam for neutron scattering, it must be made monochromatic—that is, all the neutrons but those with a specific amount of energy must be removed. Experimenters achieve this just as they produce monochromatic beams of X-rays: by reflecting the beam at a glancing angle from a crystal.

The reflection produces a beam of neutrons with specific energies in different directions; it remains to narrow the beam down by passing it through a small gap and to play it onto the target compound, which generally takes the form of a fine powder.

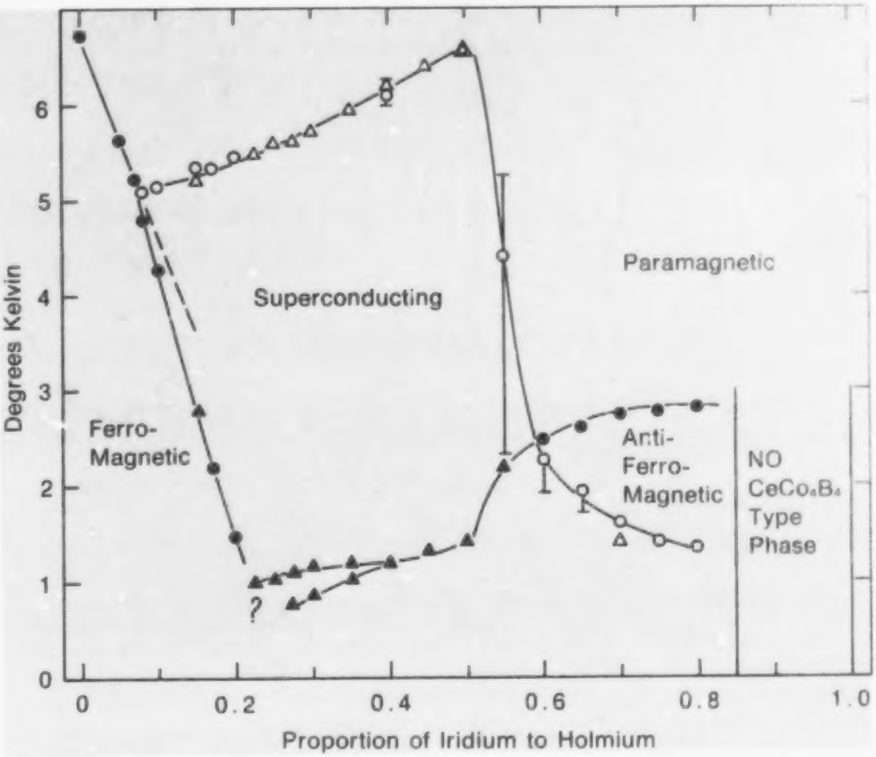
A new state

Partly as a result of neutron scattering, it became clear in the mid-to-late 1970s that superconductivity and magnetism do mix in the superconducting Chevrel phases; groups headed by Brian Maple and Oystein Fischer studied systems that showed both zero resistance to electric current and genuine, long-range antiferromagnetic order. The first part of the long search had finished. Uneasy as it might be, the alliance between superconductivity and magnetism could be made. Later on, Maple and his group at La Jolla found that superconductivity and long-range antiferromagnetic order occur in several ternary, rare-earth rhodium boride compounds made with the rare earths neodymium, samarium, and thulium.

The tougher question still remained: Could long-range ferromagnetism coexist with superconductivity? For in retrospect theorists were not surprised that antiferromagnetism could show compatibility with the superconducting state; after all, antiferromagnetic compounds possess no net magnetic field to interfere with the Cooper pairs of electrons that are essential to superconductivity. That would not be the case with a ferromagnetic compound.

The new quest quickly focused on two compounds: erbium rhodium boride and holmium molybdenum sulfide. Shortly before he died, in 1977, Bernd Matthias, along with Maple and several of their coworkers at La Jolla, found that the three-element compound erbium rhodium boride was a reentrant superconductor. It exhibits superconducting behavior when its temperature drops through 8.7 degrees above absolute zero (Kelvin), then loses it when its temperature falls below 0.9 degree Kelvin. Meanwhile, Fischer and his associate in Geneva, looking at the compound holmium molybdenum sulfide, found it also to be a reentrant superconductor; it acquired the state of zero resistance at 1.8 degrees Kelvin and lost it at 0.64 degree.

All the compounds had two things in common: First, because neither contained rare-earth impurity ions, their reentrant superconductivity owed nothing to the electron-ion interaction of the Kondo effect. Second, both were plainly ferromagnetic alloys. Finally, it seemed, physicists had in their laboratories not one but two compounds in which ferromagnetism and superconductivity coexisted.



Magnetic maze. Increasing proportions of iridium (shown along the horizontal axis), when added to the compound holmium rhodium boride, create an array of different phases as temperature increases along the verticle axis.

After M. Brian Maple

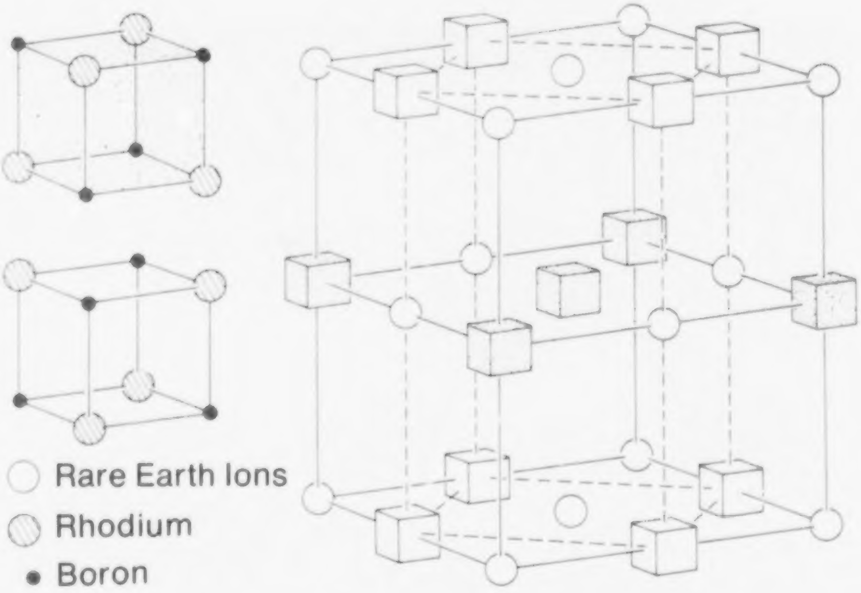
But was the ferromagnetism the genuine variety, with long-range order? Determining that required extremely subtle neutron-scattering studies. A Brookhaven group headed by David Moncton and Lawrence Passell looked at erbium rhodium boride. Another team headed by Jeffrey Lynn, using the National Bureau of Standards' reactor at Gaithersburg, Maryland, as well as the Brookhaven facility, studied the Geneva compound holmium molybdenum sulfide. Both groups detected an entirely unexpected phenomenon just above the reentrant transition temperature below which superconductivity vanished.

Moncton's group saw a peak in the neutron-scattering images corresponding to magnetic ordering with a repeat distance of about 100 angstroms. The peak first appeared at about 1.5 degrees Kelvin; it strengthened as the temperature fell until it disappeared abruptly at the reentrant transition temperature of 0.9 degree Kelvin. Below that transition temperature, ordinary ferromagnetism occurred.

Another effect complicated the picture. The phenomenon showed what physicists call hysteresis. The peak corresponding to the 100-angstrom repeat distance was weaker when induced by bringing the alloy temperature up to its reentrant temperature than by cooling the compound down.

Theorists interpret the strange experimental evidence as follows: Between 1.5 degrees and 0.9 degrees Kelvin, Matthias's erbium rhodium boride contains two separate phases, or states of matter, that coexist on a microscopic level. One is a simple ferromagnetic region that shows no superconductivity whatever; the other is a zone in which superconductivity and a type of complex and unusual magnetic order actually coexist.

The two characteristics of matter are in strong competition in such zones, however. As a result, the ferromagnetism may be slightly altered. Then, instead of lining up parallel to each other, adjacent magnetic moments actually lie with a slight angle between them; only those moments 100 angstroms apart, the repeat distance, are



Coexistence. Superconductivity and ferromagnetism can coexist in rare earth compounds with rhodium and boron. The rare-earth ions are responsible for the ferromagnetism of these compounds, while the rhodium and boron clusters contribute the pairs of electrons that create superconductivity.

M. Brian Maple

truly parallel. The arrangement represents a compromise by the magnetic component of the alloy that allows it to remain superconducting and magnetic simultaneously. "This," says Brian Maple, "is a new state of matter."

Lynn's group found a greater degree of coexistence of the two effects with the Geneva group's holmium molybdenum sulfide compound. In that alloy, no undiluted ferromagnetic zones interfere with the picture. Between 0.73 degree Kelvin and the reentrant superconducting temperature of 0.67 degree Kelvin, the compound exists entirely in a superconducting state that is magnetically ordered with a repeat distance of about 200 angstroms. As in Matthias's compound, the magnetic moments oscillate in such a way that they can accommodate the alloy's superconductivity. Strangely, the oscillatory phase does not appear when the compound is heated up through its reentrant transition temperature, as later experiments at Grenoble, France, showed.

Further evidence of the inherent compatibility has come from an Argonne group headed by David Hinks. He managed to grow a single crystal of Matthias's erbium rhodium boride. Single crystals provide more precise neutron-scattering images of their internal magnetic arrangements than do the powdered samples normally used. Studies of the crystal at the Oak Ridge National Laboratory's High Flux Isotope Reactor confirmed the coexistence in the crystal of ferromagnetic domains and the superconducting regions with an oscillatory arrangement of spins.

However they work . . .

What causes the coexistence? Why does the ferromagnetism not overpower the superconductivity rather than accommodating to it? The answer seems to be in the structure of the two alloys. "In effect," explains Maple, "you have two independent structures in these compounds." The rare-earth ions—erbium in the one case, holmium in the other—form their own crystal lattices that are largely independent of the other elements in the compounds. As a result, little interaction takes place between the rare-earth ions, which provide the alloys' magnetism, and the other atoms, which provide the paired electrons necessary for superconductivity.

However the systems work, the two alloys demonstrate that superconductivity and ferromagnetism can coexist under limited circumstances. The experimental studies have largely confirmed the theoretical predictions of two decades ago. Reentrant superconductivity exists just as Gof kov and Rusinov suggested, and the newly discovered zones of compatibility bear at least some resemblence to the cryptoferromagnetism predicted by Anderson and Suhl. Recently, another Bell Labs group—Eugene Blount and Chandra Varma—has taken a detailed theoretical look at the ternary superconductors and has predicted more precisely the lazy spiraling of magnetic moments actually seen in superconducting ferromagnetic phases. The theorists' work suggests that the observed repeat distance of 100 angstroms in erbium rhodium boride is consistent with theory and forecasts the experimentally detected first order transition from the state of coexistance between ferromagnetism and superconductivity to the simple ferromagnetic phase. Richard Ferrell of the University of Maryland has had similar success with a slightly more comprehensive theoretical approach. Both theorists and experimentalists agree that superconductivity can coexist with long-range antiferromagnetic order and an adapted form of ferromagnetism; they also hint very strongly that it cannot live with normal, long-range ferromagnetism—just as the experts thought from the beginning.

That conclusion does not mean that the field is finished. Experimentalists must now characterize more precisely the nature of the coexistent state. "The theorists do not want to say anything about it right now," says Argonne's Gopal Shenoy wryly.

Further, the physicists also face the prospect of making to order new solid phases containing the seeds of even more conflict than that between superconductivity and ferromagnetism. If the rhodium in erbium rhodium boride is partly replaced by iridium, for example, specialists predict that superconducting zones will emerge in which ferromagnetism and antiferromagnetism square off against each other. Substitution of holmium appears to produce a superconducting region in which the ferromagnetism tries to order along two different directions. So far, examinations of such compounds have only scratched the surface. As investigators look further, they will undoubtedly gain new insights into the nature of the electrical and magnetic properties of the solid state.

Does the research promise any technological dividends? At present, it is too early to say. One possibility in the minds of physicists studying superconductivity is that of creating a new series of superconductors that retain their resistance-free state to temperatures higher than the current maximum of 23 degrees Kelvin. That may eventually happen as a result of the compatibility experiments. But for now, the real fascination lies in the challenge of understanding solids in entirely new and unexpected forms. ●

The National Science Foundation contributes to the support of the work discussed in this article through its Low-Temperature Physics Program.

Letters, etc.

Chemist as hero

I want to extend my congratulations on *Mosaic* in general and the March/April 1982 [Volume 13, Number 2] issue in particular.

Mosaic is a valuable source for the communication of highly technical subjects to the general public. It succeeds in being written in plain English and also in a highly interesting style. This combination brings to the public many of the fascinating stories of research.

You are to be congratulated also on your ability to select highly capable authors with that rare talent of translating science into exciting reading for the average person.

It is my gut feeling that many Americans are interested in learning about our highly technological and scientific society. People are interested in science—but it has to be presented to the average person in a palatable way. This way, for better or for worse, might be in the form of entertainment—a TV program about a chemist "hero," for example!

Lawrence P. Verbit
Department of Chemistry
State University of New York
Binghamton

Culture gap

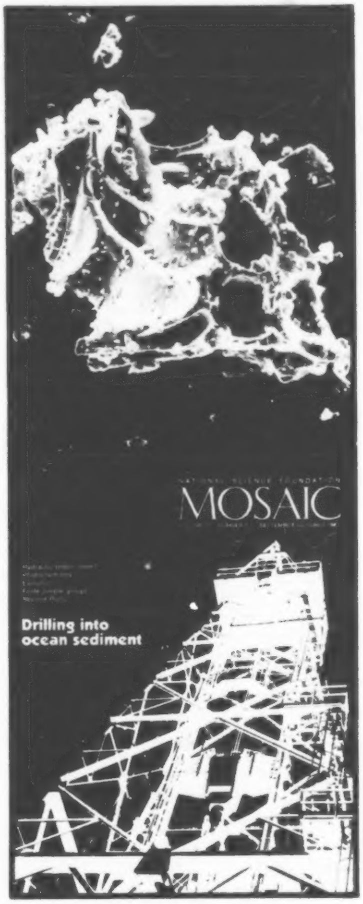
With reference to the recent article "The Roots of Science in Ancient China," [see *Mosaic*, Volume 13, Number 2] please note that the correct spelling of [Joseph] Needham's work is: *Science and Civilisation in China*. You may not care to see "civilisation" with an "s", evidently they do at the Cambridge University Press.

Russell Maeth
Center of Asian and African Studies
El Colegio de Mexico

If you care about science, you will want to read

The National Science Foundation's

MOSAIC



Discover that science can be timely, readable, *and* authentic.

"Scientists read *Mosaic* the way actors read *Variety*....I usually read it to find out about science in an area other than my specialty." That's the way one prominent oceanographer describes the usefulness of *Mosaic*. For that scientist, *Mosaic* is on target. His reason for reading the bimonthly magazine is precisely the reason it is published: to provide a perspective on the range of research areas on which the National Science Foundation's interests touch.

Mosaic is written and edited by experienced science journalists and authenticated by the researchers whose work it discusses. It can be of value to scientists who care about research areas other than their own, and to nonscientists who just care about science.

To subscribe to *Mosaic*, to send it as a serious gift to a serious friend, or to order single copies, use the order cards bound into this magazine.

National Science Foundation
Washington, D.C. 20550

Official Business
PENALTY FOR PRIVATE USE, $300

Postage and Fees Paid
National Science Foundation

U.S. MAIL

BULK RATE
U.S. POSTAGE PAID
PERMIT NO. G69
WASHINGTON, D.C.

Women's S/E Doctorates Double Between 1971 and 1981

The number of science and engineering doctorates awarded to women more than doubled between 1971 and 1981 to 4,020, or 23 percent of the total. In contrast, the number of men receiving science and engineering doctorates fell by 20 percent to 13,600. The growth in the number of these doctorates awarded to women, however, was insufficient to offset fully the decrease in the degrees awarded to men.

The 1971-81 period witnessed a dramatic shift in doctorate production trends. The number of degrees awarded in all fields reached a peak of almost 33,800 in 1973, culminating a long-term series of annual increases dating back to 1957. Between 1973 and 1978, the number of doctoral degrees awarded declined by almost 9 percent. Since 1978 annual doctorate production has stabilized at about 31,000. The trend for science and engineering degrees followed the same pattern—reaching a peak of 19,000 in 1973, declining by more than 10 percent to 17,000 in 1978, and recovering slightly to 17,600 by 1981.

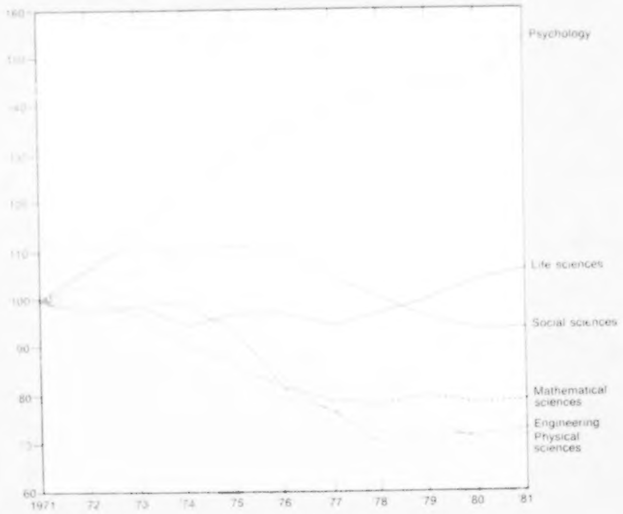
The relatively small changes between 1971 and 1981 in total and science-and-engineering degree production stemmed from rather substantial but offsetting changes in the basic factors underlying this degree production—i.e., in the pool of potential new doctorates and in the proportion of members of that pool who actually acquired these degrees. The potential pool, defined statistically as the number completing baccalaureates seven years earlier, doubled; the fraction acquiring doctorates (the completion rate) fell dramatically—by almost 50 percent.

There were substantial field differences in production trends during the 1971-81 period. Most notable were decreases in engineering (down to 2,530 or 28 percent below the 1971 level) and in the physical and mathematical sciences (down 29 percent to 3,210 and 22 percent to 960, respectively). In contrast, psychology had a dramatic 57-percent increase, reaching 3,360 in 1981. The life sciences also reached a new high level of 4,780 in 1981—a modest 5-percent increase over 1971, but this field experienced declines in some years during the 1971-81 period. The social sciences declined to 2,790, or about 7 percent below the 1971 level.

The patterns of growth and decline in the science and engineering fields were related to the changes in the sex distribution of the new doctorates. Fields with the largest proportions of women (psychology and the life and social sciences) were among those showing increases in the number of degrees awarded during the 1971-81 period; in fields with lower proportions of women (engineering and the physical and mathematical sciences) the total number of doctorates fell. Women's representation, however, increased in all science and engineering fields, but the changes were not uniform. In 1981, as in 1971, the proportion of women was greatest in psychology (44 percent in 1981) and lowest in engineering (4 percent in 1981).

The 1971-81 growth of science and engineering doctorates awarded to women is the result of both the more rapid relative growth in the female baccalaureate pool in science and engineering, which nearly tripled in those ten years, and the less rapid decline in the female doctorate completion rate in those fields. That rate fell by only 27 percent.

Index of S/E doctorate production by field, 1971-81
(1971=100)



S/E doctorates awarded by field and by sex, 1971-81

Field	1971 Total	1971 Women (percentage)	1973 Total	1973 Women (percentage)	1975 Total	1975 Women (percentage)	1977 Total	1977 Women (percentage)	1979 Total	1979 Women (percentage)	1981 Total	1981 Women (percentage)
All fields	31,870	14	33,760	18	32,950	22	31,710	25	31,220	29	31,320	32
S/E total	18,950	10	19,000	13	18,360	15	17,420	18	17,240	21	17,620	23
Engineering	3,500	*	3,360	1	3,000	2	2,640	3	2,490	2	2,530	4
Physical sciences	4,500	5	4,080	6	3,710	8	3,420	9	3,320	11	3,210	11
Math sciences	1,240	8	1,230	10	1,150	10	960	13	980	15	960	14
Life sciences	4,560	14	4,500	18	4,400	20	4,270	20	4,500	23	4,780	26
Psychology	2,140	25	2,460	29	2,750	32	2,990	36	3,090	41	3,360	44
Social sciences	3,010	13	3,370	15	3,350	20	3,140	21	2,860	26	2,790	26

*Less than 0.5 percent

Note: Detail may not add to total because of rounding.

SOURCE: National Science Foundation and
National Research Council

13/6

NATIONAL SCIENCE FOUNDATION

MOSAIC

VOLUME 13 NUMBER 6 NOVEMBER DECEMBER 1982

Free-electron lasers
Language acquisition
Gene expression
Computer graphics
Plasma reconnection
Stormscale weather

LD [redacted] C
NEWARK PUBLIC LIBRARY
FILE ROOM

Free-electron Lasers

NS1.29:

In this *Mosaic*

The relationships between the large and the small are often an inherent problem in scientific research. Animals—people among them—do not behave in groups as they do as individuals. Laboratory insights on the scale of biological molecules or cells are not readily extrapolated to the functioning of complex multicellular organisms. And in oceanography or meteorology, problems of scale are often what the research is all about.

In physics, there are theoreticians to whom the identity of the large and small, the quantum and the cosmic, the strongest force in nature and the weakest force in nature are not a difficulty; they are the whole question and the whole answer.

Experimental physicists, however, face the same problems of scale as do other scientists: The interaction of magnetic fields may be known to be intimately connected to the effects of bodies in space on each other, for example. But that doesn't make it any easier to reproduce in a laboratory reactor those relationships among fields and particles. And while inferences can often be drawn from laboratory to natural scale, conclusions reached that way can be most tentative.

As Margaret Silbar sums it up in this *Mosaic*, in her article on magnetic field reconnection: "It is not that the physicists have scaled down the cosmic disturbance to laboratory size . . . . Rather they have reproduced 'the potent effects.' . . . The parallel, then, might be said to be not quite complete." But she concludes: "Ultimately, however, it all has to come together. When that day comes, there will be only one kind of plasma physics." Or, she might have added, chemistry, or biology, or psychology, or . . . .

Warren Kornberg

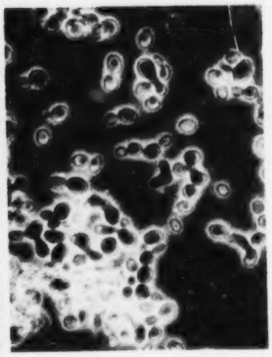
READEX FICHE best copy available

MOSAIC

VOL 13, NO. 6 NOV./DEC. 1982

2

22



29

NATIONAL SCIENCE FOUNDATION

Lewis M. Branscomb, *Chairman*
National Science Board

Edward A. Knapp, *Director*
National Science Foundation

Thomas Ubois, *Acting Director*
Office of Government and
Public Programs

Warren Kornberg, Editor
David Heffernan,
Assistant to the Editor
Joyce Latham, Copy Editor

Mosaic is published six times yearly as a source of information for the general public and the scientific and educational communities served by the National Science Foundation. The Director of the National Science Foundation has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this agency. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1984. Subscriptions are available from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription is $16.00 per year in the United States and possessions and $20.00 all other countries. Single copies are $4.25 domestic and $5.35 foreign. Send check or money order to the U.S. Government Printing Office. Communications other than those concerning paid subscriptions should be addressed to: Editor, *Mosaic*, National Science Foundation, Washington, D.C. 20550. Telephone: 202-357-9776.

Cover. The wiggler magnets being inserted into the vacuum envelope during final assembly of the free-electron laser at Los Alamos.

Los Alamos Scientific Laboratory

NEWARK PUBLIC LIBRARY
JAN 24 1983
U. S. GOVT. PUBS.
DEPOSITORY 376

Storms on a Deadly Scale *by Kendrick Frazier* 2
Mesoscale, or stormscale, weather systems are the ones that do the most damage. Predicting them has been beyond reach.

Merging Magnetic Fields *by Margaret L. Silbar* 9
In their swirling dance, magnetic fields of the earth and sun sometimes blend, or reconnect. Laboratory simulation of the action is being called an experimental tour de force.
Reconnection and magnetic monopoles 12

Lifting the Limits of Laser Frequencies *by Kendrick Frazier* 12
The free-electron laser offers a feature not available with conventional lasers: tunability across the visible spectrum.
A laser primer 17
The free-electron concept 18
Lasing at x-ray frequencies 20

Under Babel's Tower *by Lee Edson* 22
How children acquire language is a riddle for developmental linguists and the subject of debate among them. The acquisition process may be universal.

The Regulation of Gene Expression *by William Check* 29
The triggers that turn genes on or off appear to involve movement, relative position, chemical modification, and nearby proteins.

A Palette for Computers *by Allen L. Hammond* 35
Computer-generated imagery is reaching a level of sophistication that promises to challenge the most discerning eye. The power of available computers may be a limiting factor.
Computer-aided design 39

University Research Equipment Expenditures Above $350 Million in 1980 Back cover

Storms on a Deadly Scale

May 9, 1980; Tampa Bay, Florida: Just as the 35,000-ton phosphate freighter *Summit Venture* was trying to negotiate the channel spanned by the Sunshine Skyway Bridge, a sudden, violent rainsquall struck. The ship rammed the bridge supports, tearing away a thousand-foot span and killing at least 35 people. A major artery important to Florida's economy was severed; harbor operations could not return to normal for a month. Estimates are that by the time the bridge was fully repaired, as much as a billion dollars in damage may have been done to the Florida economy.

- June 12, 1975; Kennedy International Airport, New York: An Eastern Airlines Boeing 727 approaching the runway encountered a sudden, violent downdraft; such phenomena have since been termed downbursts. The plane slammed into steel light-towers, a half-mile from the runway, flipped over, and disintegrated. All 113 persons on board were killed in the accident.
- April 4, 1977; New Hope, Georgia: A Southern Airways DC-9 encountered a rapidly moving squall line and crashed as the pilot attempted an emergency landing on a highway. Sixty-three people on board and nine on the ground were killed. An investigation by the National Transportation Safety Board concluded that the heavy rain and hail through which the aircraft had flown for two and one-half minutes had caused a "total and unique loss of thrust" in both of its jet engines.
- July 19, 1977; Johnstown, Pennsylvania: Fair weather covered the United States east of the Mississippi River. The 4 P.M. forecast for southwestern Pennsylvania indicated a 30 percent chance of thunderstorms. That night meteorological processes combined to drop ten inches of rain on a small region of the Conemaugh River Basin. The resulting flash flood took 77 lives.

Every year brings storm disasters like these. They are caused by severe weather that covers only a small area and persists for a short time. To avert destruction and loss of life, forecasters would have to prepare and publicize their warnings and communities would have to respond in a matter of minutes—a difficult task.

At present, science and technology cannot anticipate weather systems at this scale and issue effective warnings. Leaders in the atmospheric science and weather service communities are beginning to believe, however, that they soon will. They also feel that the means are now at hand to move storm-warning capabilities strongly in this direction. During the next ten years, largely through a massive national storm program beginning to take shape, they expect both research and new observational and communications technology to match the task.

Storm-sized systems

The phenomena they are talking about are all small scale. They include severe squall lines, thunderstorms, tornadoes, downbursts, rain-induced flash floods, wind shears, sudden wind shifts, local heavy snows, and banded precipitation zones. Nor must they

Sturm und...

The life cycles of mature storm clouds are studied to understand the processes that produce them.

NCAR

...drang

Mesoscale, or stormscale, weather systems are the ones that do the most damage. Predicting them has been beyond reach.

by Kendrick Frazier

all be violent. They can also include freezing rains, dense ground fog, killing local frosts, and persistent temperature inversions that create local air pollution episodes. The list encompasses a good part of the weather that disrupts economic activity and affects lives and property.

Meteorologists generally refer to weather on this scale as mesoscale, or sometimes stormscale. The phenomena range from roughly one to a thousand square kilometers, or from the size of a cloud to that of several states. They fall in size between the microscale (cloud physics and diffusion and turbulence) and the synoptic scale (the overall view of the weather systems covering nations, continents, and oceans).

Mesoscale weather is far from a new concern. "It includes phenomena that have long been recognized as among the most destructive natural hazards," says Robert M. White, president of the University Corporation for Atmospheric Research. This consortium of universities is doing initial planning for the National Storm Program. "So why now, and what's new?" White asks. He believes the answer is that "there is a confluence of events which enables us to make major progress in the understanding of mesoscale phenomena; new technology of observation is now available, our theoretical understanding enables us to formulate a meaningful program, and the national need is clear."

The scars of violent winds: utility poles snapped by a mesoscale storm in Boulder, Colorado.

Wide World Photos, by permission.

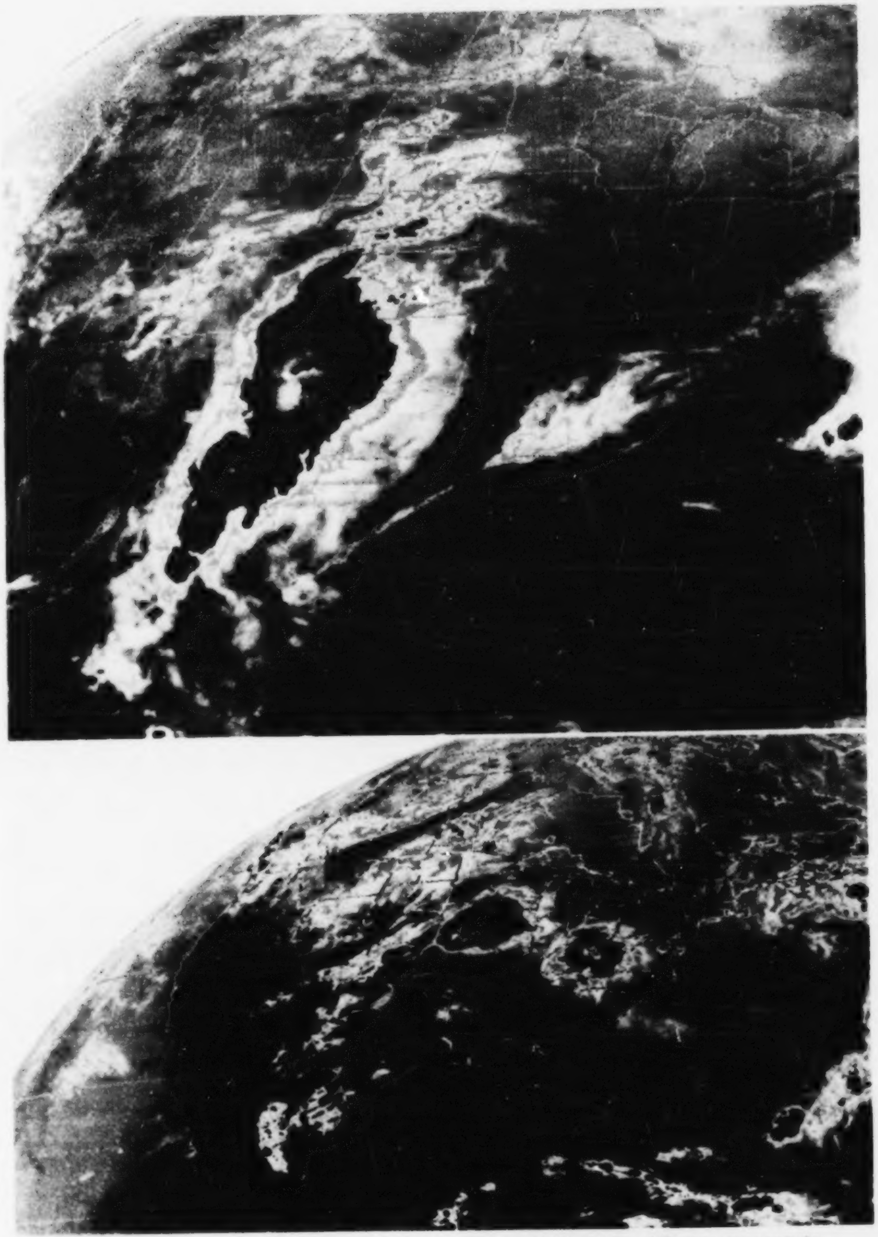
Why now? What's new?

Two areas of advance account for the sense of opportunity now felt by atmospheric scientists. One is the advent of new observational and communications technologies: research Doppler radar; satellite-borne sounding instruments that provide quantitative meteorological data, not just imagery, from space; ground-based profiler systems that can use microwave scanners; and other remote-sensing technology for complementary, ground-based probing of the lower atmosphere. The next generation of radar planned by the National Weather Service will combine Doppler radar's storm-profiling abilities with computer-aided analysis and visual display, keeping in step with new communications technologies.

The second area of advance is the increased understanding that has come through use of numerical models to simulate mesoscale weather phenomena. Models are in use in many parts of the world to produce the regular forecasts of large-scale weather. But localized, mesoscale models are much more difficult to use as predictive tools. Advances have been made, however, to the point where Roger Pielke of Colorado State University, a specialist in numerical modeling on these scales, can call it "a powerful theoretical tool."

George Benton, of Johns Hopkins University, sees a natural progression in the new emphasis on the mesoscale. Benton is chairman of a steering committee planning a national storm research program.

Beginning with the first distribution of weather observations by telegraph in the 1840s, through development of the frontal-



Satellite images. Infrared satellite images taken in June 1982 of two different mesoscale storm complexes show them to be large, circular cloud shields. Such systems are probably responsible for beneficial summer rain as well as violent storms.

Robert A. Maddox/NOAA

system concept in the early part of this century, and then on to the Global Atmospheric Research Program in the 1960s and 1970s, meteorologists have gained a fairly comprehensive understanding of the large-scale circulation of the atmosphere and the behavior of large-scale weather systems. "We now understand the physics of large-scale circulation fairly well," Benton says. "That was a necessary first step."

Interaction between scales is important, too. J. Michael Fritsch of Pennsylvania State University, another meteorologist active in planning the National Storm Program, emphasizes this problem. For example, mesoscale conditions can modify the synoptic environment in such a way that the likelihood of a tornado decreases but the chance of a flash flood increases.

"If we are to forecast [meso]scale events, we need to account for many interacting scales of physical processes and include mesoscale memory," says Fritsch. He expresses awe at the fact that "we are on the verge of doing precisely that." He also agrees with Benton, White, and others that the time is now right for a full-scale effort.

Benton says he first had the idea for a major program in mesoscale meteorology in 1969. "It fell with a resounding thud, and I now think quite rightly. We were too early." Doppler radar was not available, and neither was temperature sounding from satellites. Similarly, Benton observes, "our ability to simulate the small-scale events of the atmosphere numerically has moved ahead very rapidly; it is now far ahead of where we were ten years ago." There has been a revolution in the observing technology and in numerical modeling, "and we might as well put it to use," he adds.

Meteorologists agree that there is a serious lack of basic understanding of organized precipitation systems in the mesoscale domain. Yet intriguing recent discoveries emphasized a common tendency for clouds to organize on the mesoscale, suggesting that fundamental factors may be involved. For example, there is the work done at the National Oceanic and Atmospheric Administration's Environmental Research Laboratories by Robert A. Maddox. He has identified and defined a particular class of weather system: the mesoscale convective complexes.

Cold beginnings

Maddox identified these complexes by analyzing recently available and enhanced infrared imagery from geosynchronous meteorological satellites—the same ones that provide photos of North American weather to daily newspaper and television forecasts. This type of satellite-data display has shades of black, gray, and white to show progressively lower cloud temperature. These are some of the first routine quantitative meteorological data from satellite imagery.

Mesoscale convective complexes, Maddox found, begin with nearly circular, continuous shields of cold cloud tops on a scale that is huge by mesoscale standards. By his definition, the area of cold—cloud tops at 32 degrees below zero centigrade or lower—exceeds 100,000 square kilometers. Maddox believes the cloud-top convection complexes "are a class of convective weather system heretofore unrecognized." They pose many new puzzles. The phenomena and effects surrounding them are not forecast by existing numerical models of the atmo-

---

Frazier is the author of The Violent Face of Nature *and* Our Turbulent Sun.

sphere. Nor, says Maddox, is their physics understood.

Mesoscale convective complexes are common over the central United States in the summer. They can persist for 12 hours or more and are probably responsible for much of the widespread and beneficial summer rain in the central part of the nation. "It is possible," says Maddox," that these systems dominate the precipitation and convection climatologies for the growing season over the United States wheat and corn belts."

Linear cells

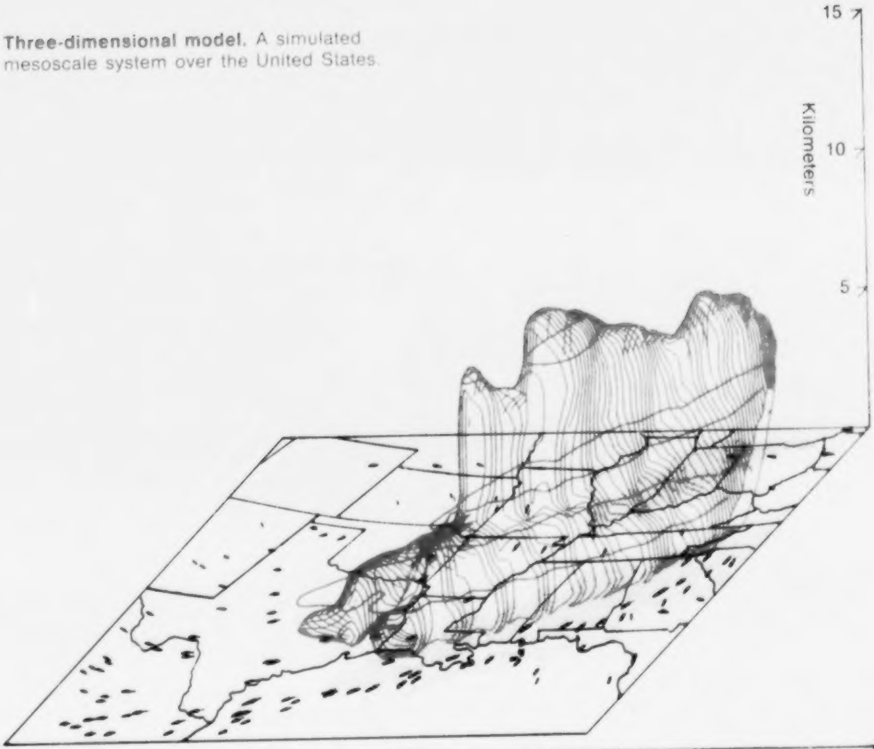
Another pattern recently revealed is the mesoscale structure of rainfall within large Pacific cyclones that move onshore in the northwestern United States. The regions of relatively heavy precipitation as these storms move onshore, reports Peter V. Hobbs of the University of Washington, are organized into large, linear, mesoscale rainbands. They come in several kinds (warm-frontal, warm-sector, cold-frontal both wide and narrow, prefrontal cold surge, wavelike and postfrontal), and they can cover areas of a thousand to ten thousand square kilometers. The rainbands themselves are composed of smaller mesoscale areas, or cores, of heavier precipitation roughly ten to a hundred square kilometers in area.

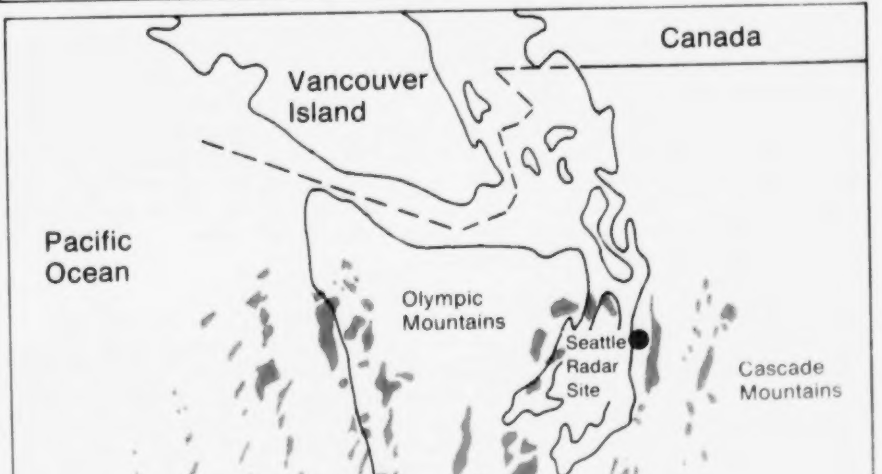
The mechanisms responsible for the organization of precipitation into rainbands are not understood, but Hobbs says the rainbands are inherent features of mature cyclones over the Pacific Ocean. As they move over the land, topographic effects may become important. Small hills, for example, may help trigger or enhance rainbands in cyclonic storms in regions where the air is unstable. But in the case of narrow cold-frontal rainbands, which are anchored to a cold front, large mountain barriers can interfere with the low-level, southerly flow of air. This flow provides the front with its main source of moisture and can disrupt the precipitation.

Recently, Hobbs and his colleagues have managed to elucidate a fine-scale structure in narrow cold-frontal rainbands. They made use of the National Center for Atmospheric Research's portable CP-3 Doppler weather radar, probing the cold fronts associated with cyclones approaching the coast of Washington State from the Pacific Ocean. Doppler radar can trace the position of a cold front accurately by locating the wind-shift line associated with it. Reflectivity data can track regions of precipitation with high resolution.

It had been thought that these narrow (about five kilometers wide) rainbands of heaviest precipitation, which typically strad-

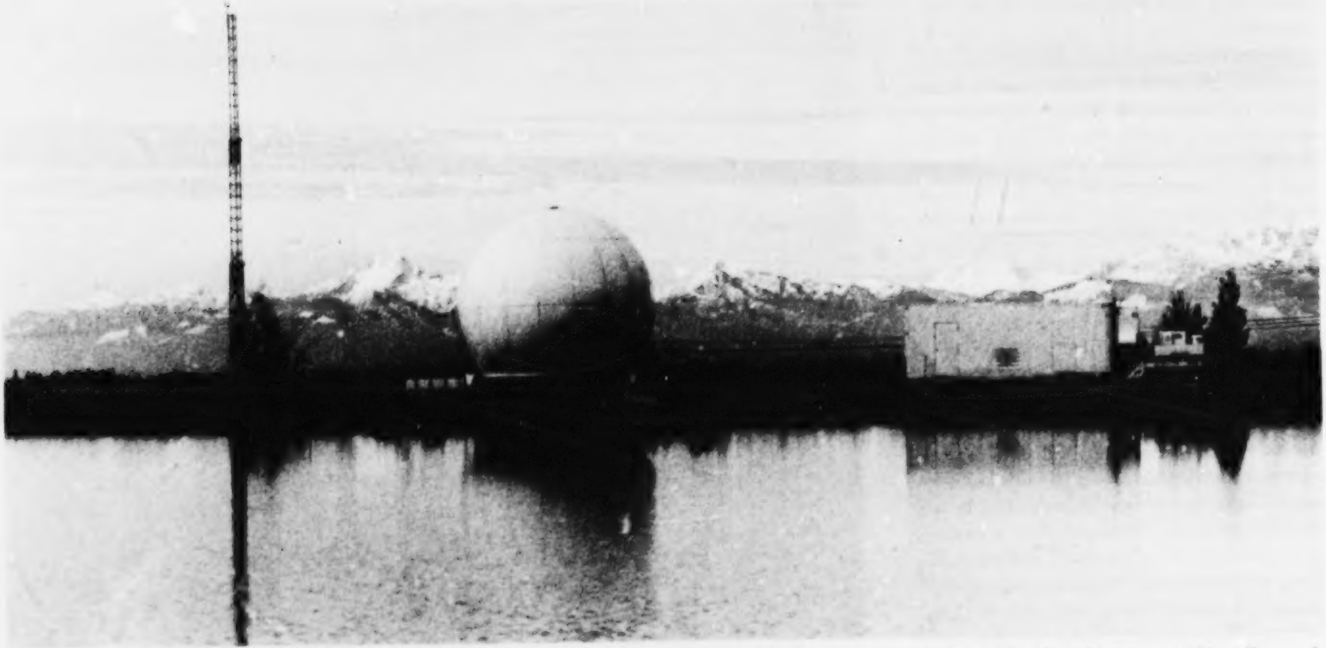
Three-dimensional model. A simulated mesoscale system over the United States.





Precipitation cores. Radar probes show precipitation cores associated with a narrow cold-frontal rainband as it passed from west to east over the Pacific Northwest.

P. V. Hobbs



Studying cyclonic storms. A scanning Doppler radar is housed in the large dome dominating a field site used by the University of Washington.

dle surface cold fronts, had essentially no cellular structure. The University of Washington scientists have found otherwise: Rather than consisting simply of lines of convection, the narrow cold-frontal rainbands exhibit a remarkable and uniform substructure. When the band is viewed in detail by Doppler radar, its heaviest precipitation is seen to be organized into regularly spaced, small mesoscale elements roughly elliptical in shape. The long axes of the ellipses are essentially parallel to each other and oriented at an angle of about 35 degrees to the front.

The dimensions, spacing, and orientation to the front of these elliptical precipitation cores are often remarkably similar in different storms. In six different instances observed, the precipitation cores were typically four kilometers wide by about seventeen kilometers long. The average distance between cores was four kilometers. The angles between the long axes of precipitation and the front varied only from 32 degrees to 38 degrees. The cores are relatively low-lying features (most of the rainfall begins below three kilometers), and they remain on the front while moving progressively along its length. Such information is by and large descriptive. "A quantitative explanation for the small mesoscale organization of precipitation on cold fronts," says Hobbs, "would appear to be a challenging problem."

With the addition of still another research tool, numerical modeling, Robert B. Wilhelmson of the University of Illinois has been studying the relationship between the vertical structure of atmospheric conditions and the type and behavior of the storms that develop. His numerical modeling work is at the small end of the mesoscale domain, up to about 250 kilometers. This ranges "from the scale of a tornado to the scale of line squalls," as he puts it.

Quantifying storms

Wilhelmson and his colleagues have been able to simulate various classes of storms, but the one most studied to date is the kind that produces tornadoes, often called a tornadic thunderstorm. In studying such phenomena, the mesoscale model developed by Wilhelmson and Joseph B. Klemp of the National Center for Atmospheric Research is typically 100 kilometers across and 16 kilometers high. Such information as winds and rain content is computed at grid points one kilometer apart horizontally and half that vertically.

Wilhelmson says that such work has shown ways that the environment of a tornadic storm can be distinguished from that of other kinds of storms. This is done, he says, by appropriately relating the vertical changes in wind direction and speed to height changes in the temperature and moisture. And, Wilhelmson adds, Klemp and Morris Weisman of NCAR have shown that "we can define a number that's characteristic for tornadic thunderstorms." Usually, such storms develop when horizontal winds change in speed rapidly with height, when there is plentiful low-level moisture near the ground, and when the temperature decreases rapidly with height in the system, Wilhelmson explains.

One recent study combined Peter Ray's Doppler radar analyses at the National Severe Storms Laboratory with a three-dimensional numerical storm simulation. The subject examined was the detailed structure of a tornadic storm that hit near Del City in central Oklahoma. The computer-simulated structure of the mature storm matched closely that of the actual storm.

The researchers found that the storm exhibited certain important features essential to maintaining its longevity and promoting its transition to the tornadic phase. For instance, the rotational character of the storm separated the precipitation from the updraft, so the disturbance did not quench its own source of heat and moisture. The storm also oriented the resulting downdrafts so that low-level convergence of moist air along the ground-level gust front was reinforced, also helping to maintain the storm.

Split storms

Wilhelmson and Klemp have also used their three-dimensional mesoscale model to explain the observed splitting of certain severe thunderstorm radar echoes into two parts. The model simulates self-sustaining right- and left-moving storms that arise through splitting of the original. The simulated right-moving activity develops a structure strongly resembling the accepted conceptual model of a right-moving thunderstorm, which often spawns tornadoes and destructive hail after the split.

The Illinois and NCAR scientists have recently applied this kind of analysis to a split storm in Oklahoma and found qualitative similarities between the model and the actual event. They caution that this kind of work is not yet usable for forecasting, but it can help to create a better understanding of storm development and structure.

Many other experimental efforts to develop mesoscale numerical modeling have recently begun. Some have shown remarkable abilities to simulate individual mesoscale circulations. Atmospheric scientists are encouraged, although they agree broadly that it will be a long time before they can provide any useful support for actual forecasts of mesoscale weather.

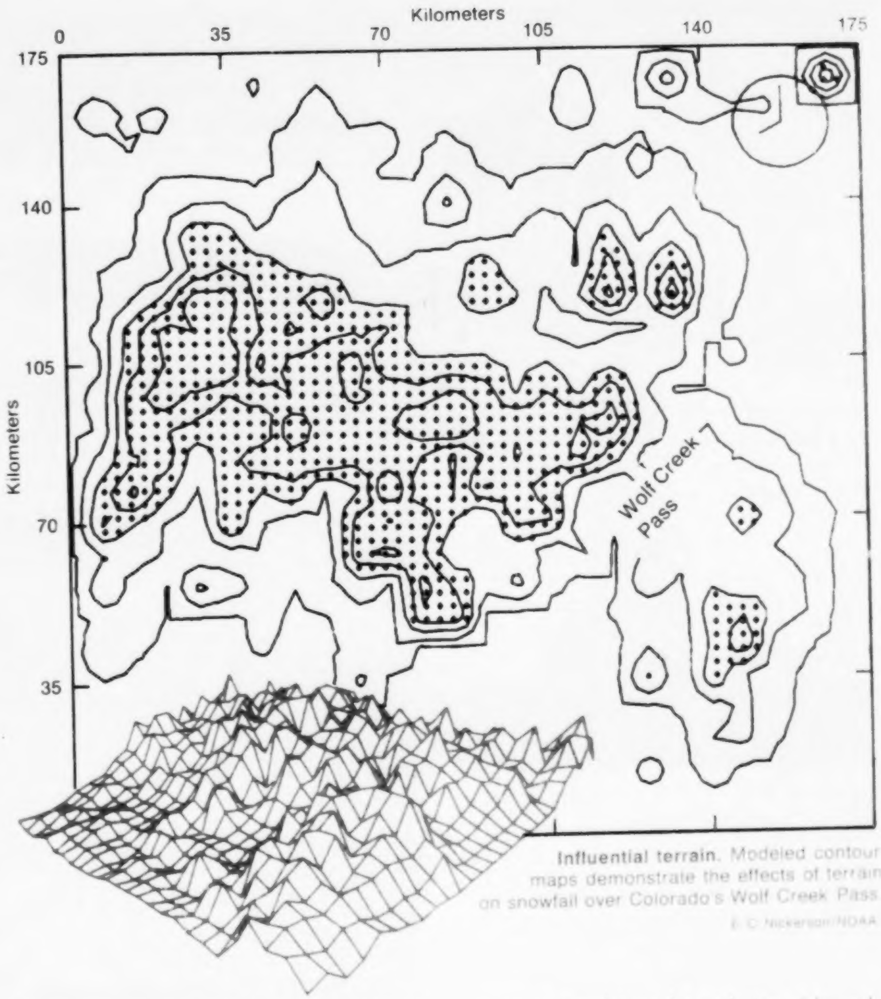
Terrain effects

Roger Pielke of Colorado State University has been using mesoscale numerical modeling to study weather systems influenced by terrain effects. Terrain influences have been notoriously difficult to model correctly and many modelers have shied away from them. But they are crucial to modeling many important kinds of weather systems, including lake-effect snowstorms, sea and land breezes, mountain valley flows, and urban effects, many of which are extremely important in air pollution studies. One of Pielke's efforts shows how a sea breeze can cause a mass of industrial air pollution from a source on the coast to head out to sea at upper levels but then recirculate at low levels back onshore.

Modeling of terrain-induced mesoscale weather has an important advantage over other, externally influenced weather, Pielke explains. The terrain remains constant; its influence does not change. "Terrain-induced mesoscale systems offer one of the best opportunities to improve weather forecasts as well as to prepare more accurate environmental assessments," he notes.

Everett C. Nickerson of NOAA's laboratories in Boulder has applied still another three-dimensional mesoscale model to the prediction of precipitation over mountains. One study focused on the Alsace Plain on the French-German border. The model did a good job of predicting both the amount and location of precipitation over the Black Forest mountains to the east, says Nickerson, but it did not do as well over the plains.

Another study, applied to the island of Hawaii, did a good job of predicting precipitation, even identifying a small desert area on the mostly rainy, windward side of the island, adds Nickerson. And a model of the mountainous Wolf Creek Pass area of southwestern Colorado shows, perhaps not surprisingly, that the terrain has a great influence on the large amounts of snowfall there. "The model does a pretty good job of giving the horizontal distribution of precipitation," says Nickerson.



Influential terrain. Modeled contour maps demonstrate the effects of terrain on snowfall over Colorado's Wolf Creek Pass.

E. C. Nickerson/NOAA.

Nickerson believes such efforts will eventually lead to better forecasting of mesoscale weather. "I believe that the outlook for improvement in local precipitation forecasts is very encouraging, especially where terrain is a forcing function." But numerical models of mesoscale phenomena are seen primarily as a research tool, a source of basic understanding of mesoscale weather. Many researchers caution that it would be unrealistic to expect them to help weather forecasts very soon.

Dynamical numerical models have shown tremendous progress during the past ten years in simulating the evolution of hazard-producing storms on small scales," says Carl W. Kreitzberg of Drexel University. There have also been many important advances in observation, communication, and display systems. The obvious question is, adds Kreitzberg, "Can these advances be combined to provide the very useful predictions of localized hazardous weather our society needs so badly? The answer at the moment is no."

Modelers too often reside in the "nice neat world" of simulated data and assumed conditions, says Christopher M. Hayden of NOAA's National Environmental Satellite Service Development Laboratory in Madison, Wisconsin. "Now enters the leprous misfit of the world of observation."

Observing technologies

Nevertheless Hayden is not pessimistic. "I think the situation is right and the future bright for improved numerical mesoscale forecasts," he notes. One reason for

On watch. An NCAR meteorologist coordinates the operations of several Doppler radars. The console display screens show the most violent portions of the thunderstorms.

NCAR

his optimism is the new observational technologies becoming available.

The vertical atmospheric sounder, an experimental infrared-and-visible scanning radiometer that is on board the geosynchronous satellite GOES-5, has the ability to monitor temperatures and moisture at every level down through the atmosphere at a horizontal resolution of 14 kilometers. At the same time, a ground-based profiler, a combination of microwave radiometer, wind-profiling radar, and surface sensors, can continuously monitor atmospheric winds, temperature, and humidity from the ground up. The profiler has been undergoing detailed tests north of Denver and has been performing well.

"It is already doing somewhat better than was predicted in accuracy," says C. Gordon Little of NOAA's Wave Propagation Laboratory in Boulder. "I am frankly surprised at how fast this research tool has produced data of this quality."

The surface-based profiler and the satellite-based sounder complement each other nicely, Little points out. The satellite does not measure winds well; the profiler does. The satellite is especially good at measuring the upper atmosphere; the profiler is very good in the lower atmosphere.

Atmospheric scientists say the profiler offers a useful adjunct to—or even a replacement for—the existing network of rawindsondes. These are the instrumented weather balloons launched twice daily from weather stations spaced approximately 400 kilometers apart and are the main source of weather data for current forecasts. The 400-kilometer grid spacing, however, is too coarse to pick up many mesoscale conditions.

And then there's what is called NEXRAD, the planned next-generation of weather radar expected to be ready in the late 1980s or early 1990s. It will replace today's aging, conventional radars with Doppler radars that can provide measurements of motions within storms. Frederick P. Ostby, director of the National Severe Storms Forecast Center in Kansas City, believes it would be "the single most important contribution to an improved warning system."

Combine profiler with vertical sounder and then add NEXRAD, declares George Benton, and scientists and forecasters will have a set of powerful new tools at their disposal for better anticipating stormscale weather. "Somewhere in the 1990s—I'd guess by 1995," he adds, "I think we'll have a new generation of observing systems in meteorology able to measure the temperature, winds, and humidity of the atmosphere entirely by remote-sensing—by ground and by satellite."

Benton is convinced that the new observational technologies, expected advances in mesoscale modeling, the host of new communications and display systems to convey weather warnings to the public, and a concentrated push in mesoscale research and forecasting will all eventually pay off. The results, if he and his colleagues are right, will be faster, more accurate, and more timely forecasts of local precipitation and the hazardous weather on the scale that affects the public most.

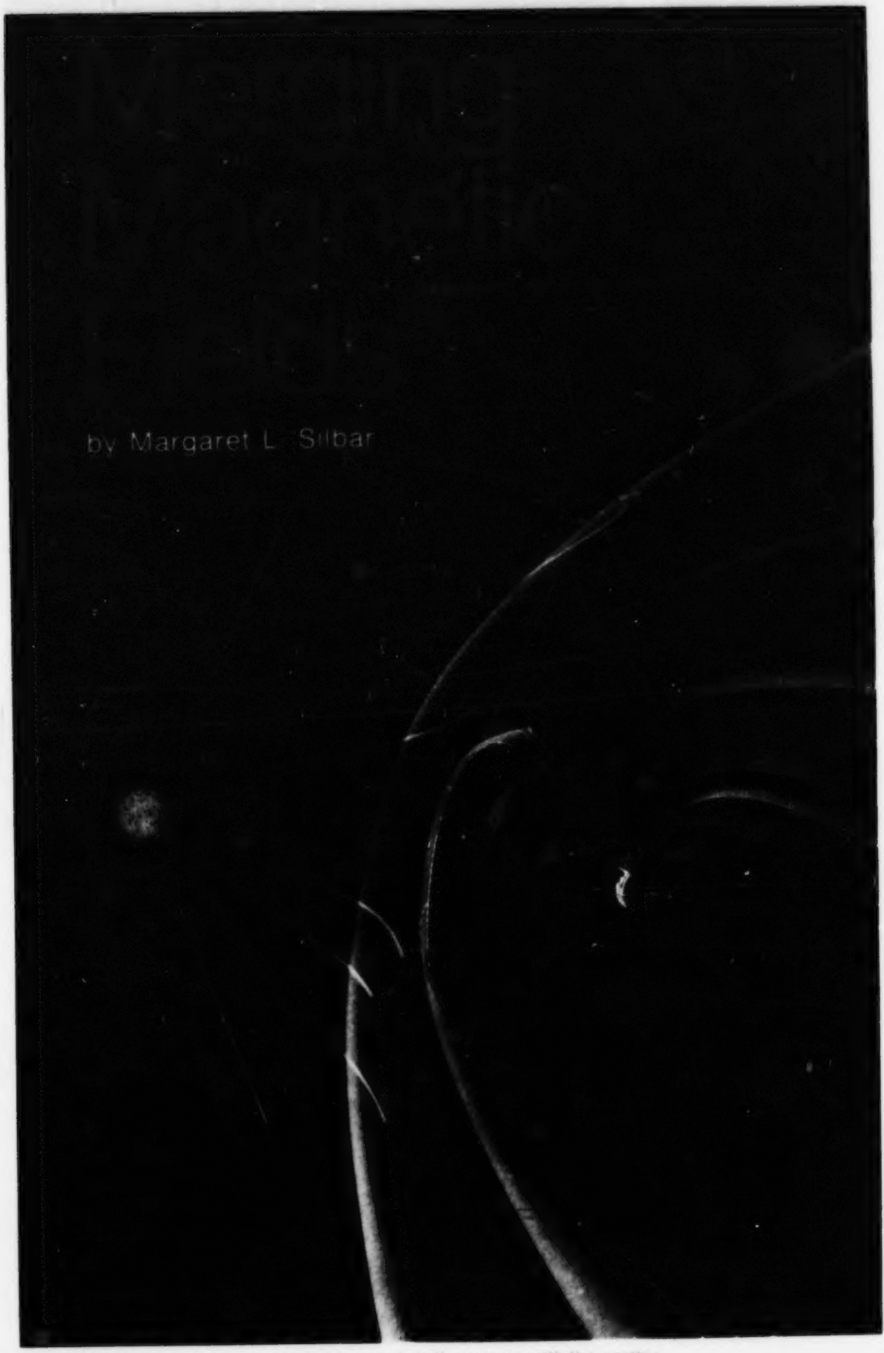
"The time is now right," says Benton. "The technology is available or on the horizon. Scientific methods are at hand, and the incentive exists."

"Barring catastrophe," says NOAA's William Hooke, "I believe the textbooks in 50 years will say that there were enormous strides made in national weather services in the 1980s and 1990s." •

The National Science Foundation supports the research discussed in this article through its Meteorology and Experimental Meteorology Programs and through support of the National Center for Atmospheric Research.

In their swirling dance, magnetic fields of the earth and sun sometimes blend, or reconnect. Laboratory simulation of the action is being called an experimental tour de force.

by Margaret L. Silbar



Connection. The sun's magnetic field occasionally merges with the earth's.

NASA

The earth is nestled in a magnetic cocoon, buffeted by a steady wind of the charged solar particles that stream toward it along the sun's magnetic field lines. Ordinarily the earth's magnetosphere shields the planet from these products of the violent magnetic events that occur on the solar surface. But now and then a loop of the sun's magnetic field boils up toward the earth and bursts through the planet's cocoon like a bolt of magnetic lightning. In the process the sun's magnetic field merges with that of the earth.

Such mergings are by no means unique to the earth. They occur around other magnetized planets as the solar wind swirls by. They also occur in the deep interior of the sun and other stars, around black holes, and in the centers of galaxies. Such a merging can be expected to take place wherever there are magnetic fields to interact with the plasmas, thin gases, of electrically charged particles. Since magnetic fields and plasmas are extremely common throughout the universe, that means practically everywhere. (See "When North Becomes South," *Mosaic*, Volume 13, Number 5.)

Space probes in the neighborhood of the earth have made possible the study of many of the gross details of this plasma-magnetic field interaction. But the microscopic details have been, as Robert Rosner of the Harvard Smithsonian Center for Astrophysics puts it, "hidden in a black box." A few scientists have managed to pry the box open a bit and peek inside, only to have the cover slam shut again; they could get two magnetic fields to merge, but not consistently.

Recently, however, in a laboratory experiment at the University of California at Los Angeles, physicists have found a way to prop open the lid of the box and photograph the events inside. The motion pictures they made of the merging process are "oh-so-beautiful to watch, like the waves in the ocean," says UCLA's Reiner Stenzel who along with Walter Gekelman succeeded where the others had failed.

An old problem

The idea of connecting two stiff and separate magnetic fields, together with their associated plasmas, dates back to the thesis work of Ronald G. Giovanelli in Canberra, Australia. Giovanelli, who managed to get his Ph.D. in observational astronomy during World War II, studied solar flares, the sudden brightenings on the solar surface. As was known at the time, these always seem to occur in conjunction with a group of sunspots.

After a while Giovanelli came to the conclusion that the explosive burst of energy that characterizes a flare comes from the intense magnetic field of a sunspot. Sunspots occur in magnetic pairs, in which a magnetic field points out of one spot on the sun's surface and into the other. It seemed to Giovanelli that flares tended to occur along lines separating pairs of sunspots.

Perhaps, Giovanelli suggested, magnetic energy is somehow released along such neutral lines, where the oppositely directed magnetic fields cancel. Indeed, it is now thought that this release of energy gives a big kick to the charged particles in the solar wind; the flare may release as much magnetic energy in a minute or two as the sun ordinarily emits as light in the course of a year.

The noted English astrophysicist Fred Hoyle was the external examiner for Giovanelli's thesis. "I'm reasonably sure that's how he got involved," says James W. Dungey of Imperial College in London, then a student of Hoyle's. (In his own Ph.D. thesis in 1950, Dungey was the first to discuss the problem of the merging of magnetic fields, sometimes called reconnection.) In any case, Hoyle saw both "the promise and the problems." He proposed to Dungey that he consider how such a process might work.

From the sun. Coronagraph shows a flare ejecting solar material.

NCAR

Dungey says he spent about two years "seeming to get about nowhere." In fact, he recalls, he was riding his bicycle on the way to tell Hoyle he had "a bum problem" when "it occurred to me what Hoyle would say—I knew Hoyle would declare, 'Why take any notice of Lenz's Law? It might not apply!' I promptly did a U-turn right outside Woolworth's in Cambridge and went back to Maxwell's equations."

A loophole

The rule known as Lenz's Law is a component of the four equations with which James Clerk Maxwell succinctly summarized all electric and magnetic phenomena, making them one. As Dungey discovered, laying a base for many later studies, a loophole in that rule allows the magnetized solar wind to interact with the magnetized earth. "In general," explains Hans Griem of the University of Maryland, "plasmas resist change; a plasma is almost like a superconductor." Thin layers of moving charged particles, between plasmas from different sources, work to keep the plasmas separated.

Moreover, many plasmas can be considered, to a very good approximation, collisionless. The solar wind, spiraling outwards from the solar corona like the spray from a spinning lawn sprinkler, is an example. In its trip across the 150 million kilometers between the sun and earth, the streaming plasma is so thin that, on the average, any electron in it will collide with only one other electron—effectively no collisions at all.

One way to cross the boundary separating the solar wind from the earthly plasma, the Van Allen belt, is to pry one of the solar wind's electrons from the solar magnetic field line around which it spirals. It isn't easy. Every electron tends to stick with the field line on which it starts out. Yet it can be done; a solar-wind electron can diffuse into the earth's magnetic field. Plasma particles are sometimes knocked off their field lines by collisions with other particles or by some other kind of instability. And as Maryland's James Drake says, "Little instabilities tend to grow into big ones."

In the earth-sun case, these collisions allow the separate magnetic fields to merge across the thin, elusive layers of plasma at the nose and tail of the magnetosphere. What might be considered a magnetic slingshot is created, converting magnetic energy into the kinetic energy of the plasma. Great numbers of electrons then are shot down into the atmosphere. Their subsequent collisions with the atmospheric atoms can produce auroral displays and other terrestrial effects.

Such flows of energetic electrons were first seen in the 1960s, in the tail of the magnetosphere, with detectors on board the *Vela* satellites. Similar but indirect evidence for merging at the magnetosphere's nose has only recently been found. What exactly is happening at the microscopic level remains unclear. It was to probe the geometry of these thin plasma layers and the merging process associated with them that Stenzel and Gekelman began designing their laboratory plasma more than three years ago. Helping almost from the beginning was one graduate student, Norbert Wild. Two others, David Whelan and Manuel Urrutia, later joined the group.

Reducing scale

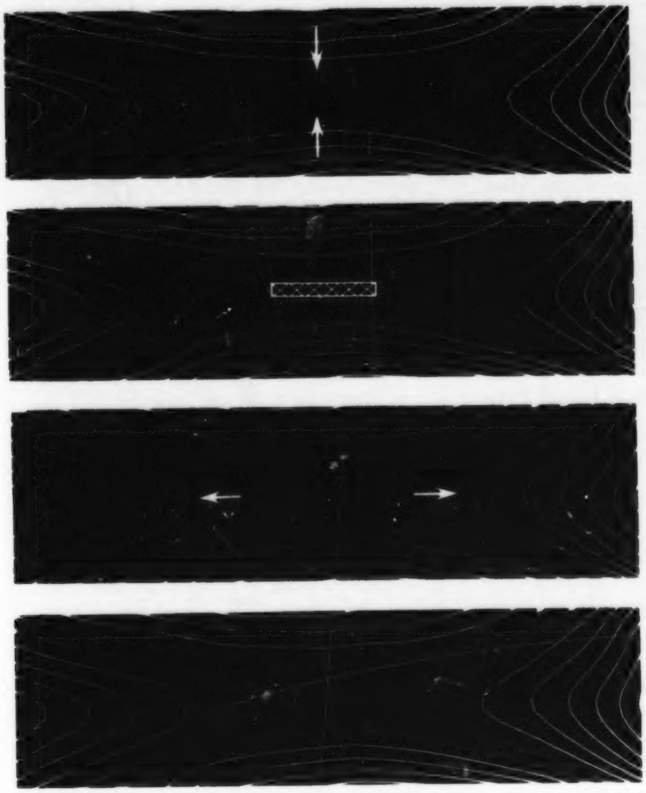
"It took a lot of scrounging before it all came together," Gekelman recalls. High-current cable came from government surplus, diodes from a friend, a large variable transformer from the back of the lab (which some years before had housed a cyclotron). The researchers were offered—free!—a large vacuum chamber, provided Gekelman went to northern California, took it apart, and packed it up for shipping. He went.

The real problem, however, was the plasma itself. Since no container can withstand the heat needed to strip the electrons from the atoms of a gas, the trick in making a plasma is to apply the heat internally. There are various ways of proceeding, but the UCLA team decided to use an electrical discharge from a cathode. A cathode is essentially a charged piece of metal that, when heated, boils electrons from its surface. Coating the surface of the metal with special oxides allows the boiling to take place at lower, more easily attainable temperatures.

The cathode yields energetic electrons. These collide with the atoms in the gas surrounding the cathode and knock other electrons from their atomic orbits. These then join the flow in a cascade that becomes an avalanche. As more electrons are freed, they collide with more atoms, freeing still more electrons. After a short while, the original gas of neutral atoms has become a plasma of electrons and ions.

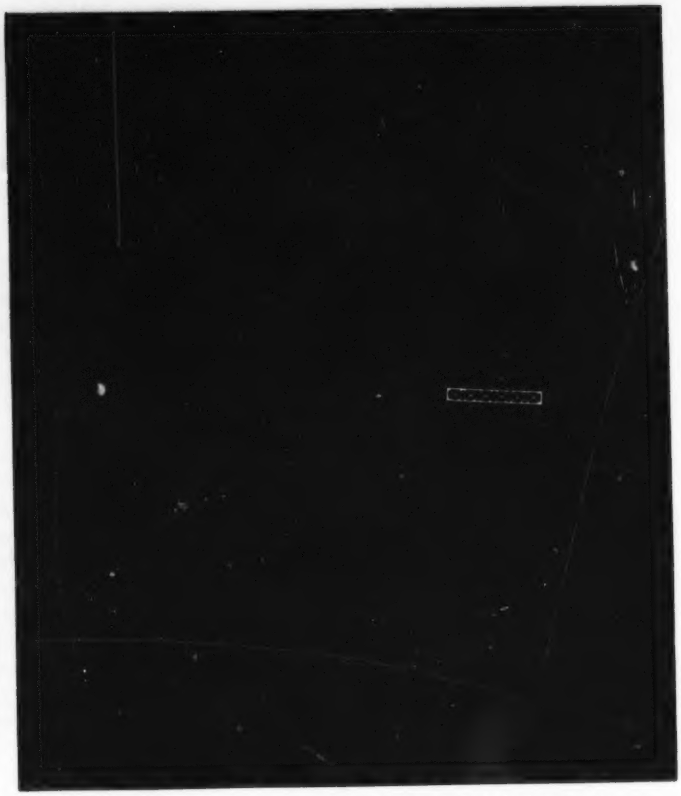
The resultant plasma keeps its shape, courtesy of a longitudinal magnetic field provided by external coils, but it by no means lives forever. It tends to cool quickly as the ions and electrons recombine into ordinary neutral gas atoms. The plasma, consequently, must be renewed continuously by energetic electrons from the cathode.

The UCLA physicists wanted a plasma large enough to reveal the kind of structure they believed was hidden within. Ultimately, the



Simulated reconnection. In four steps (from top down, 0.5 microsecond apart), the innermost lines of magnetic force move vertically toward each other, touch at the null point and separate horizonally. The process repeats for the next innermost pair.

After Reiner Stenzel, Walter Gekelman



Touching loops. The oppositely-directed magnetic field loops of the sun and earth occasionally touch and merge across thin layers of plasma. As a consequence, electrons plunge into the earth's atmosphere, producing terrestrial effects like the aurora.

Margaret Silbar

volume of the plasma is dependent on the size and efficiency of the cathode. Commercially made cathodes tend to be small, of the order of ten centimeters across or less. The UCLA physicists therefore developed what is perhaps the world's largest oxide-coated cathode. It is a meter in diameter and is housed in a blue cylinder a meter and a half in diameter and two meters long.

The first time the California team turned on the cathode, it all but melted. Now, however, the argon plasma they are using is as stable as a vacuum tube. One hallmark of its stability is that one can insert measurement probes into it. In contrast, if one tries to put probes into a fusion-device plasma (see "Two kinds of plasma physics," accompanying this article), Gekelman says, "either the plasma or the probe is lost."

Observing the merging

Once set up, Stenzel and Gekelman spent two years trying to understand the structure of the magnetic field and the properties of the plasma. It is, in fact, the careful probing—diagnosing—that has led colleagues to label the work an experimental tour de force; earlier experiments of other groups were not as finely tuned.

It takes about as much computer backup to run their experiments, observes Stenzel, as a space probe requires. The plasma conditions at 300 points in each of several transverse planes are measured at thousands of time steps. Each such snapshot records the direction and strength of the magnetic field. It is as if the physicists were measuring the direction and number of iron filings in the field around a bar magnet.

With the help of a computer, all these different field vectors, as they are called, are stored and sorted out. Gekelman recalls that he and his colleagues tried in the beginning to record the lining up of vectors by hand. To do this for just one plane took a hundred hours, a task a computer can do in a fraction of a second.

This made them wonder how best to present the data being collected. "Even for us," Stenzel says, "it is sometimes very complex." Deciding that "a moving picture is worth a million numbers," they made their first movie. "Having looked at the magnetic field and stored its entire history," Stenzel says, "one sees what happens in time."

Indeed, in the movie one can watch magnetic field lines, as depicted by the computer program, flowing together. Each line can be thought of as a contour, the sum of the individual vectors. The lines change as the magnetic field strength changes. In some critical frames, the externally applied magnetic field at top and bottom can be seen to increase. This forces the field lines to move inward, closer to the region where there is no magnetic field.

Then the two field lines, which were initially separated, begin to touch. One connects with its oppositely directed partner. At the next moment, two newly formed field lines become separated. They move off to the right and left of what has become the null region, a place where there is no magnetic field at all, toward the regions of higher magnetic field strength.

The merging process is repeated until all the excess energy in the field is depleted. The newly merged field lines, as they move

---

Silbar, a science writer, just completed a year working at the National Bureau of Standards.

Reconnection and magnetic monopoles

A moving electrical charge or current generates a magnetic field, an object as real as bread and cheese. Fields exist at every point in space, and they vary smoothly in going from one place to another. For many purposes, however, it is useful to visualize the magnetic field in terms of its field lines.

It is along these lines, and tracing them, that iron filings line up in their familiar semicircular and symmetrical loops on a sheet of paper held over a bar magnet. But as can be seen from the pattern of the iron filings, the lines of this magnetic field never end or begin at a single point. Rather they form loops without crossing, creating the appearance at the ends of the magnet of being smeared. Lines of an electric field, by contrast, do start and end on discrete electric charges.

The reason magnetic field lines behave as they do is that there do not seem to be any free magnetic charges that are analogs of the electron. Or at least not many. That is, the universe generally lacks point particles (so-called magnetic monopoles) to which the magnetic field lines can anchor. When James Clerk Maxwell explained in 1864 that electricity and magnetism were real facets of each other, the possibility of an isolated magnetic charge simply did not occur to him. Seventy years later, P.A.M. Dirac suggested that there may be a quantum mechanical symmetry between electricity and magnetism. Just as electric charge exists in discrete amounts—multiples of the electron charge—so too must magnetic charge.

"It would be surprising," Dirac concluded at the time, "if nature made no use of the magnetic monopole." Exhaustive searches for conjectured monopoles have probed everything from mica 248 million years old to sludge from the ocean's bottom. Despite years of effort, none were found. (See "Magnetic Monopole," *Mosaic*, Volume 12, Number 3.)

This leads to the point that magnetic fields would not be as prevalent as they are in the universe if there were an abundance of magnetic monopoles. This was probably first noted by the University of Chicago's Eugene Parker. Monopoles would, by accelerating to ever higher energies, devour the cosmic magnetic field, he proposed.

There may, however, be a few monopoles around. In fact, one may have announced itself on Valentine's Day, 1982, to Blas Carbrera, a scientist at Stanford University.

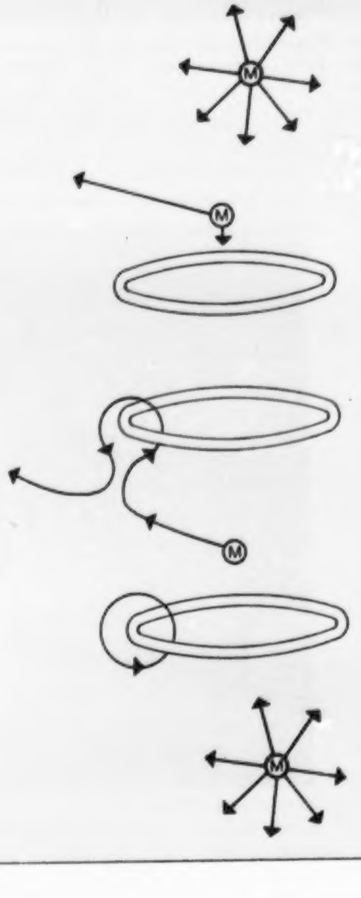
The possibility hinges on a special feature of a superconductor: It repulses magnetic field lines. Ordinarily, magnetism and superconductivity cannot coexist. (See "Superconductivity and Magnetism: An Uneasy Alliance," *Mosaic*, Volume 13, Number 5.) In this property a superconductor resembles a collisionless plasma: Both repulse the magnetic field of another body or plasma.

Carbrera has been carefully observing a superconducting coil of niobium encased in a thin lead shell. His idea is that once a year or so on average a primordial magnetic monopole—perhaps as massive as a bacterium—would meánder through the hole in his superconducting toroid. This way of looking for a magnetic monopole had never been tried before.

The idea of the experiment is as follows: When the monopole is still far away, its magnetic field lines would, like the electric field lines of an electron, radiate outward from the particle. As the heavy particle passes through the niobium coil, most of its magnetic field lines would be stretched and dragged far behind it; they cannot pass through the superconducting material.

Eventually, when the monopole is far below the superconducting loop, its magnetic field must resume its everyday shape, with field lines radiating spherically from it. This requires tearing off those parts of the field lines that have been held by the superconducting loop.

The field lines held behind would be twisted in such a way that, at the tear point, the field goes to zero; when the snap finally occurs, it produces closed magnetic field lines encircling the superconducting material. This induces a predictable step in the circulating electric current in the ring. Carbrera saw this and it led him to postulate the passing of a monopole. •



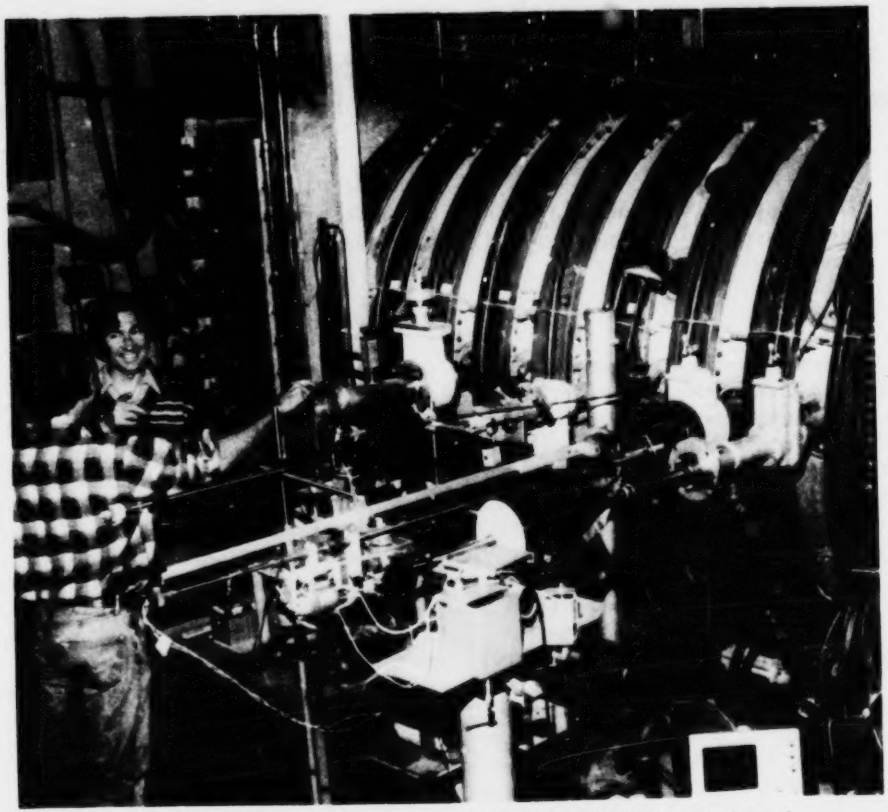
Calling card. A magnetic monopole, its field lines radiating outward,

passes through a superconducting ring.

The field snags on the superconducting material, through which it cannot pass.

The monopole tears loose, leaving a portion of its field around the ring, inducing a step in the ring's circulating current. It was such a step that led Stanford University's Blas Cabrera to propose that a monopole might have been detected for the first time.

Margaret Silber.



Generating plasmas. Reiner Stenzel (facing camera) and Norbert Wild at the housing of the cathode that is the heart of Stenzel's and Walter Gekelman's laboratory reconnection

right and left, eject jets of plasma. (In the earth's magnetotail, this plasma moves both toward and away from the earth.)

Two important points became clear from the initial experiments. First, merging is not necessarily a slow and steady process; an applied magnetic field can be made to diffuse rapidly into the plasma. This is not what would be expected if the plasma were initially acting in a truly collisionless fashion. Exactly how imperfect the plasma is, however, is a subject of study. Second, as physicists have long surmised, merging in a plasma is quite different from that in a vacuum. In a vacuum it occurs at a point, while in a plasma—as in the UCLA experiment and in the sun-earth interaction—it occurs over what is, in effect, a long, neutral sheet.

New directions

The impulsive, somewhat random geomagnetic storms seen on earth when the sun flares up have long been a practical as well as scientific problem. Such storms not only produce auroras but also disrupt communications. Television signals can start out from, say, New York and end up in Mexico. The energetic particles released in the flare would also be a danger to an astronaut lacking appropriate shielding. Such particles can easily pierce an aluminum suit. They are even being considered a possible threat to pilots of the space shuttle.

Two years ago, Stenzel says, he and his colleagues were not in a position to contribute much to an understanding of geomagnetic storms. Modeling these erratic storms had to wait until physicists learned how to produce a geometry like that thought to exist in the earth's magnetotail, the elongated portion of the magnetosphere away from the sun. The length of the tail is a measure of how much energy is being stored. The longer it is, the more energy it contains. The tail can stretch out to a thousand times the earth's radius. As the magnetic field lines in the tail are stretched and the energy is being stored, a neutral sheet is formed in the tail. This sheet separates two regions with oppositely directed field lines.

Ferdinand Coroniti, a theorist at UCLA, calls the sheet "an unnatural state, an unstable state." It breaks up in what is basically a merging process, like that in the movie-frame record produced by Stenzel and Gekelman. In this case, however, the two oppositely directed fields do not, as in the UCLA film, come together at only one point. Instead, they touch at many different ones. Closed field lines, new loops of magnetic field, form and clump into tube-like regions that float away from each other.

Such tearing of the magnetic field line also happens, destructively, in experimental nuclear fusion devices like the doughnut-shaped Tokamaks. In fact, the observation of the effect in magnetically contained plasmas in a Tokamak is one piece of rather convincing evidence that reconnection does indeed occur. In a Tokamak, the plasma is at first magnetically confined to the toroidal, or doughnut, shape of the device. Externally applied magnetic fields keep electrons circling on their own field lines. Then, through reconnection, islands start to form as field lines merge. With their formation, the once nicely directed field lines start to wander, eventually hitting a wall of the device. The plasma then is lost.

Tearing is also believed to lead to electrical disruptions in nature. These are the fingerprints of magnetic storms, and these are what the UCLA physicists have now set out to study. They have pulsed a new electric current through their experimental plasma, letting it grow in small increments. At a certain critical value, they find it can no longer carry any more of this current.

"It is not just limiting the current, but it disrupts it. The plasma breaks down," Stenzel says. That has catastrophic effects. In magnetic storms there is said to be a large release of electrical energy over a very short distance. And this, says Stenzel, "is precisely what we saw in our last experiment."

The UCLA experiment provides a model of large-scale geophysical phenomena. It is not that the physicists have scaled down the cosmic disturbance to laboratory size; that would be a hard problem indeed, says Stenzel. Rather they have reproduced "the potent effects." In the sun-earth case, the magnetic fields are not very strong at all by laboratory standards, but the distance between the two is immense. By contrast, in a laboratory experiment, the device and the distances are small while the strength of the magnetic field is large. The parallel, then, might be said to be not quite complete. But until there are microscopic studies of the real magnetic-field interactions of the sun and earth, the extrapolation from models, with all its inherent weaknesses, will have to do. •

The National Science Foundation supports the research discussed in these articles through its Divisions of Physics, Atmospheric Sciences, Material Research, and Astronomical Sciences.

LIFTING THE LIMITS O

by Kendrick Frazier

A laser is a device for the creation, amplification, and transmission of a narrow, intense beam of coherent light. Just as few now remember that the acronym radar derived from the words radio detecting and ranging, the word laser no longer seems to require its derivation: light amplification by stimulated emission of radiation. Laser, short and onomatopoeic, seems the perfect word for an intense source of light of photon densities millions of times those at the surface of the sun, yet narrow and controllable enough for surgery inside the human eye.

Within three years of the demonstration of the first laser in the summer of 1960 by physicist Theodore H. Maiman—based on theoretical work by Charles H. Townes, Arthur L. Schawlow, and others back to Albert Einstein—a laser was threatening to bisect James Bond in the movie *Goldfinger*. That is how fast the laser became part of our popular and technological consciousness.

Since then, lasers have been used to measure the distance to the moon, monitor slippage across the San Andreas fault, drill holes in diamonds, track troop movements at night, and guide bombs. They balance gyroscopes, inscribe microelectronic circuits, separate chemical isotopes, implode deuterium pellets. And they read the prices encoded in Universal Product Code lines on packages at the supermarket checkout counter. Clearly, the laser is an intrinsic part of modern life, science, technology, and commerce. (See "A laser primer," accompanying this article.)

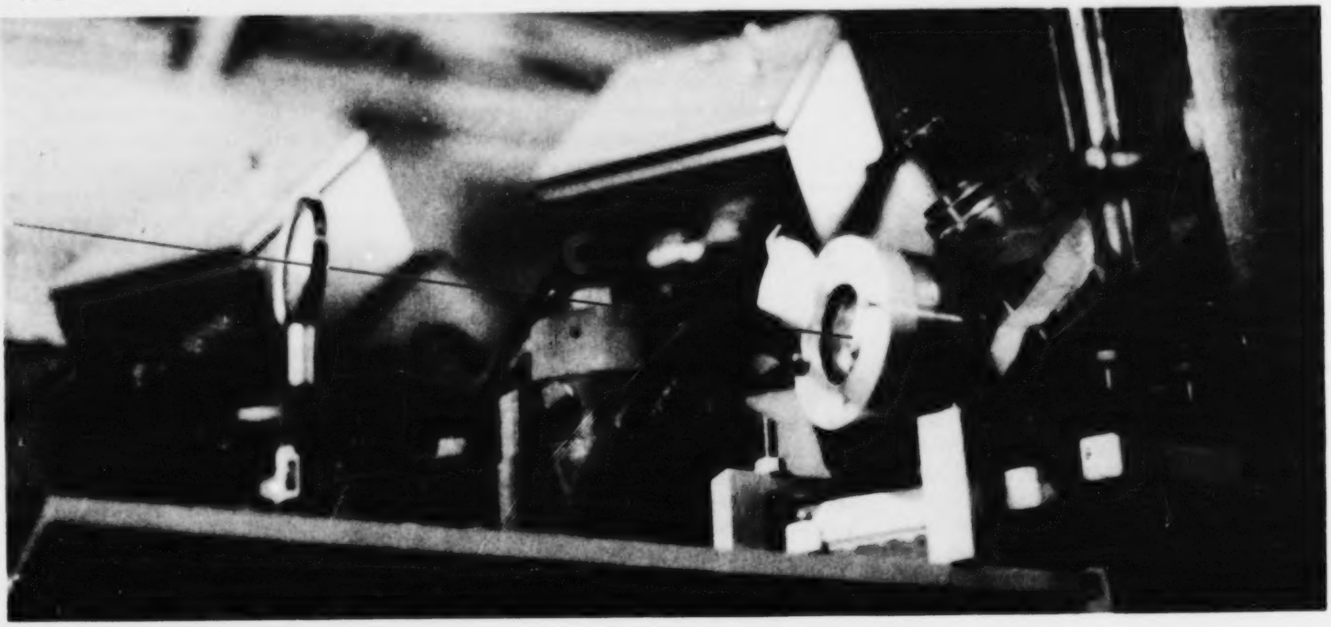
But for all their variety and versatility, conventional lasers suffer inherent shortcomings. Most, for instance, operate at only a few fixed frequencies. This is comparable to the limits of a radio receiver with a few unchangeable frequency selection buttons and no other tuning control.

For some laser applications, that lack of flexibility does not matter. But for others it does: Laser spectroscopists need light of the precise energy (frequency) of a chemical bond under study. Moreover, for lasers of high average power, the buildup of heat is a troublesome limitation; to industrial-scale chemical processing, the cost of laser operation is an important concern; and for many operations, more energy-efficient lasers are needed.

For all these reasons a revolutionary laser that substitutes free, unattached electrons for ordinary solid or gaseous lasing materials is provoking intense interest within the laser-science community. The free-electron laser offers to make laser light available at any and every frequency from the far infrared to the far ultraviolet. Now it is available at just a few frequencies here and there in the near-infrared and visible-light portions of the spectrum.

This laser makes available an enormous supermarket of frequencies ranging from roughly ten gigahertz in the longer, millimeter-wave portion of the spectrum to ten

Widely tunable. The free-electron laser developed in Orsay, France, was a collaborative effort between United States and French scientists. (The laser beam is artifically enhanced.)



LASER FREQUENCIES

The free-electron laser offers a feature not available with conventional lasers: tunability across the visible spectrum.

billion gigahertz in the far ultraviolet at the short-wavelength end. (A gigahertz is a billion—10^9—cycles a second.) All the visible portion of the spectrum is in there too, of course—that small slice centered around one million gigahertz.

The free-electron laser can span this range because it is radically different from all other kinds of lasers; it does not depend on stimulated emission from excited atoms or molecules. In fact, there is no molecular or atomic lasing medium—no ruby, no carbon dioxide, no semiconductor crystal. The free-electron laser taps the energy of free, or unattached, electrons produced in electron accelerators and moving at relativistic velocities—a very significant portion of the speed of light. They do require accelerators of a million to hundreds of millions of electron volts, however, and for many applications they are not the kind of device every chemist is likely to have around the laboratory.

Early suspicions

The origin of the free-electron-laser concept is generally credited to physicist John Madey of Stanford University. "When I was a grad student at Caltech in the mid-sixties," he recalls, "lasers were pretty new. We were trying to understand how they worked." The concept of an excited atom being stimulated to emit radiation seemed simple enough, but it was not clear that an atom was really necessary for stimulated emission. Bell Labs scientist Dietrich Marcuse and M. V. Federov of the Soviet Union had outlined a process for inducing radiation by having charged particles scatter an electron beam, although this was at radio rather than optical frequencies.

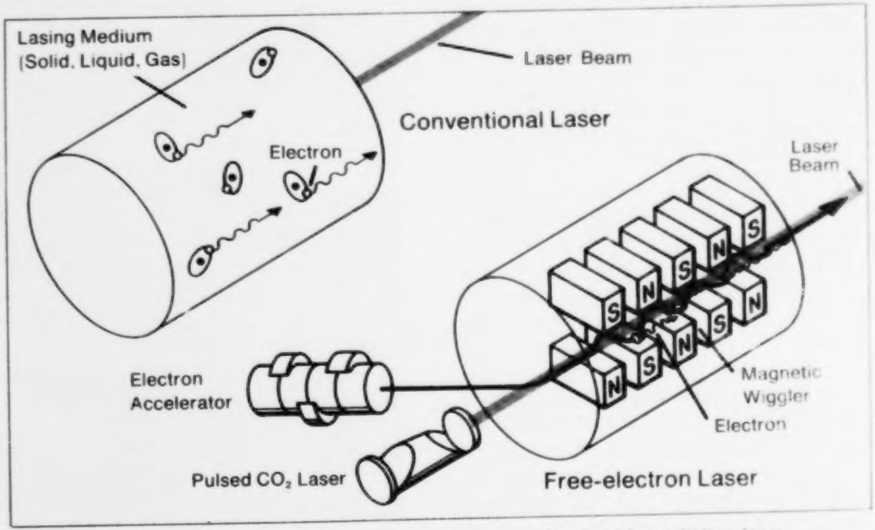
Madey was able to move the notion forward when he ran into an article by Richard Feynmann, the Caltech physicist. Feynmann "had a proof," recalls Madey, "that any radiation process in which photons were emitted could give stimulated emission. Inevitably, there would be an enhancement in the rate of radiation if some initial radiation was present. That meant we were not necessarily confined to atoms or molecules."

For five years, off and on, Madey worked on radiation processes involving interaction of photons and electrons. Eventually he conceived the notion of having a high-energy electron pass through a static but periodically alternating magnetic field, giving up photons—often at light energies—in the process. The magnets would come to be called wigglers. (See "The free-electron concept," accompanying this article.)

In 1971 Madey published in the *Journal of Applied Physics* the paper outlining the concept of the free-electron laser: "This paper deals with the radiation emitted by a relativistic electron beam moving through a periodic transverse dc magnetic field," it began. Finite gain (amplification), it went on, would be available under the appropriate conditions from the far-infrared through the visible region. It raised the possibility of laser-type amplifiers and oscillators at these frequencies, with the further possibility of partially coherent radiation sources in the ultraviolet and X-ray regions. "Not only was there interesting physics, but it actually looked as though there might be a potentially useful device," Madey recalls. "You not only [had] optical gain and no intervening media but also a continuously tunable device."

An experimental program got under way in 1972. Four years later the free-electron laser had been developed, and in April 1977, Madey and five colleagues from Stanford's High Energy Physics Laboratory (David Deacon, Luis Elias, C. J. Ramian, H. A. Schwettman, and T. I. Smith) announced their success in *Physical Review Letters*: the first operation of a free-electron laser. They noted that the usual laser constraints associated with atomic structure would not apply to a laser based on stimulated radiation by free electrons. "In this letter we report the first operation of a free-electron laser oscillator," the scientists said.

This first free-electron laser used the Stanford University superconducting linear accelerator to fire an electron beam of 43 million electron volts through a helical magnet whose polarity alternated every 3.2 centimeters.



Conventional vs. free-electron. A conventional laser amplifies light by radiating from electrons vibrating around the atomic nuclei of the lasing medium. A free-electron laser amplifies light by radiating from electrons wiggling through an alternating magnetic field.

Los Alamos National Laboratory

Radiation passing through the helix along with the electron beam was amplified, and mirrors at each end of the interaction region fed the light back through the system to be amplified again and again.

The laser light produced had a wavelength of 3.4 microns (in the infrared), a result of the particular electron energy chosen. It had an average power of 0.36 watts and a peak power of 7,000 watts. The power radiated by the oscillator was roughly 100 million times that of the spontaneous, or unamplified, radiation. The Stanford group had clearly demonstrated the capability of a free-electron laser to operate, as they described it, "at high power at an arbitrary wavelength."

"That experiment proved conclusively that free-electron lasers really work," says Charles A. Brau of Los Alamos National Laboratory. He heads the Free-Electron Laser Project there. "It stimulated the great interest in these devices that exists at present."

Wide-open range

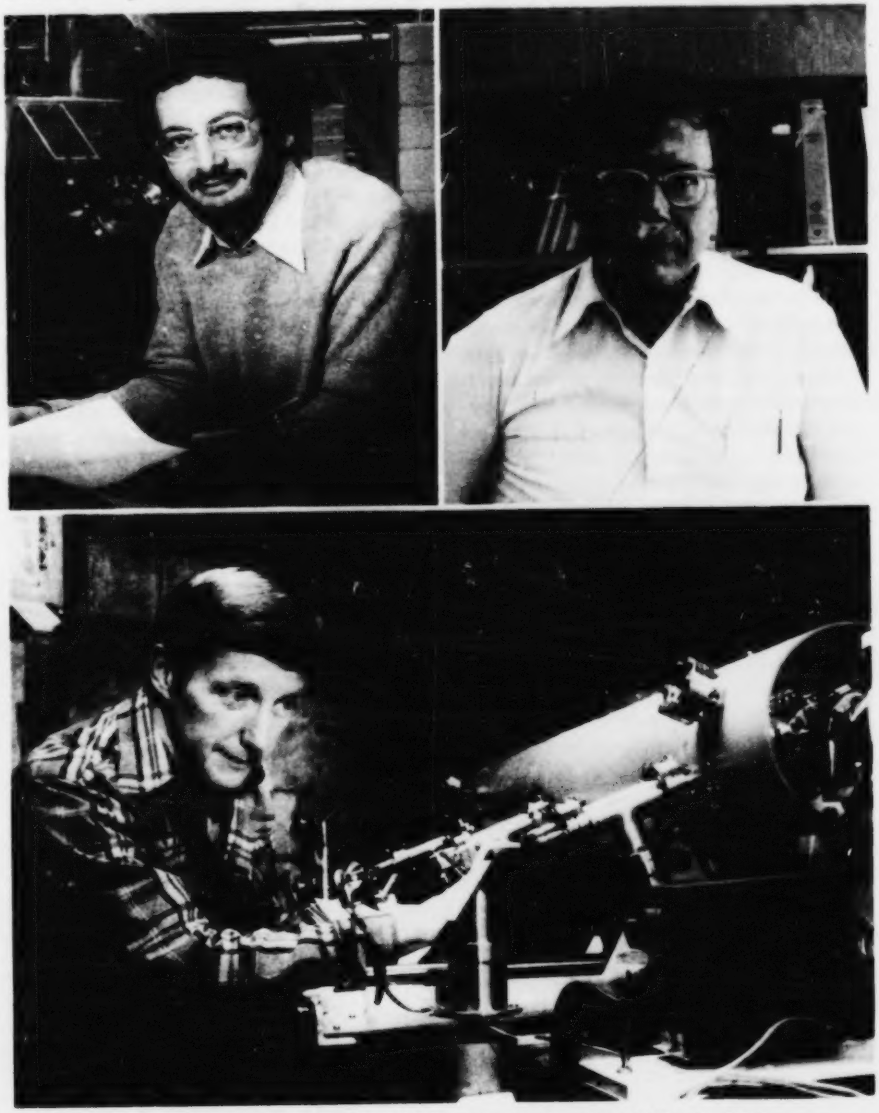
Since no atoms are being excited in a free-electron laser, and the frequency of laser light produced does not depend upon the atomic structure of any substance, that frequency can be readily changed by varying either of two quantities: the spacing between the magnets in the wiggler array or the energy of the electrons in the beam. The first way takes advantage of the fact that the distance from positive to negative and back again to positive polarity along the magnet array specifies the wiggler's wavelength. The shorter the wiggler distances the faster the electron is made to oscillate and the higher the frequency (shorter the wavelength) of the light it emits. The second way to change the frequency of the emitted light exploits the fact that the greater the energy (velocity) of the beam, the higher the frequency.

What gives the free-electron laser such extraordinary wavelength-tuning flexibility is what Bell Labs scientist C.K.N. Patel calls the "magic parameter." It is an essentially simple equation that makes the wavelength of the laser light produced approximately equal to the wavelength of the wiggler magnet divided by the square of the electron's energy. Thus, with a wiggler magnet of 10-centimeter periodicity and electrons of an energy of 20 million electron volts, the laser light produced will be at an approximate wavelength of 10/400 (0.025) centimeters, or 250 microns.

That is in the long infrared wavelengths. To get a wavelength one-fourth as long, either the electron energy can be doubled or the magnet spacing diminished to produce wiggler periodicity only a fourth as great. Free-electron lasers can be linked to more powerful accelerators to produce higher energies and shorter wavelengths, all the way up into far-ultraviolet and even near-X-ray regions. The electron accelerator used can range from the size of a desk top (for the long, far-infrared wavelengths) to the size of a room to the size of a building (for the shortest wavelengths).

Charles Brau's experimental program at Los Alamos is one of the nearly dozen under way in the United States and abroad. To him the challenge of designing and demonstrating efficient free-electron lasers is an important goal in itself, but he believes applications for chemical processing, where tunability and cost are the important factors, look especially promising. It could make economic sense, he says, to produce all but the cheapest, commonest chemicals in this manner. "Free-electron lasers," says Brau, "may revolutionize lasers and their applications."

"Historically, every new source of radiation has led to a significant increase in the types of interaction that can be investigated," declares a forthcoming report on free-electron lasers for the National Research Council by a subcommittee of its Solid-State Sciences Committee. "The demonstration of a new



Investigators. Among those doing work on the free-electron laser are (clockwise from upper left) Claudio Pelligrini of the Brooknaven National Laboratory, John Madey of Stanford University, and Charles A. Brau with the free-electron laser at Los Alamos.

Brookhaven; Kendrick Frazier; Los Alamos.

source of radiation, the free-electron laser, promises to be one such significant advance."

In the subcommittee's view, six characteristics make the free-electron laser special: tunability, high average power, potentially stable power and frequency, very short pulses with peak powers, coherence, and narrow bandwidth. The combination of these characteristics, it says, makes possible studies of condensed-matter physics; work in the spectroscopy of atoms, molecules, and ions; and surface studies in the presence of absorbed molecular species (as in catalysts)—"to name just a few."

The report predicts that many other scientific investigations will become possible as scientists become more aware of the capabilities of free-electron laser radiation. "If the promise of the current projected capabilities is realized, extensive novel scientific opportunities will become available." And that is just in solid-state science.

In Brau's view, welding and metalworking, in addition to chemical processing, look very promising "at least superficially." And, he says, "laser fusion, power transmission to satellites, laser propulsion for maneuvering in space, and even transmission of power from solar space stations back to earth are other [speculative] applications." An announcement from Los Alamos in 1981 reporting on Brau's latest results also mentioned possible military applications such as "strategic communications with submarines and directed-energy weapons for antisatellite and antimissile uses."

Nearer term

These kinds of projected applications, which require high-power lasers, are admittedly a long way off. "Surely it will be a long time before free-electron lasers make a

A laser primer

Electrons in atoms can occupy different energy states. In the normal situation, the electrons are in the ground or lowest-energy state. When the atom is excited—when energy is added to it by collisions with other atoms or by radiation—the electrons jump to an excited state. From the excited state, an electron will soon drop back down to the ground state. In doing so, it will emit a photon (a packet of electromagnetic energy) of an energy corresponding precisely to the difference between the electron's excited and ground states. The process is called spontaneous emission.

In 1917, Einstein showed another way an atom can interact with radiation. What happens, he wondered, if an atom, already in an excited state, encounters a photon? He showed that in this circumstance, the photon induces the electron to jump from the excited state to the ground state, still emitting a photon in the process.

So where at first there was only one photon, now there are two, the original one and the emitted one. This process is called stimulated emission. It happens only when the atom is already in an excited state and the energy of the incoming photon equals exactly the difference between the atom's excited and ground levels.

The photon emitted by the excited atom will also be of that frequency, will be in phase with the first photon, and will move with it in the same direction as well. This means two photons of identical energy and frequency moving in lock step, as it were, with the peaks and valleys of their wave forms matching each other exactly.

It was this principle of stimulated emission that was used in developing the laser concept in the 1950s. In a laser, the atoms or molecules in a substance used as the lasing medium (a solid, gas, or liquid) are excited by some means (for instance, a flash lamp emitting powerful conventional light). This causes a large proportion of the atoms or molecules temporarily to jump to higher energy states.

Now, if a photon of light of the right frequency strikes one of the excited atoms, the atom will emit another photon of the same frequency, and those two photons will strike two more atoms, resulting in four photons, which will strike four other atoms, resulting in eight photons, and so on. In the simplest case, all the photons are exactly of the same frequency, in phase, and traveling in the same direction. This produces a sudden burst of coherent radiation, as all the atoms discharge in a rapid chain reaction. Lasers are usually built so that their emitted light is reflected between mirrors at opposite ends of a resonant cavity, building up the intensity with each pass. One end of the cavity is designed to be less than totally reflective; through it passes the intense, highly focused beam of coherent light called a laser beam.

The wavelength of this light depends upon the transition energies of the atoms or molecules in the substance used as the lasing medium. That first laser, built at the Hughes Research Laboratories in 1960, used a synthetic ruby crystal as the lasing substance. A ruby is made up of chromium atoms dispersed evenly within a host crystal of aluminum oxide. In these chromium atoms, the energy difference between the excited state and the ground state corresponds to a wavelength of 6,943 angstroms, the wavelength of the deep red light that ruby lasers emit.

In a helium-neon gas laser, a high-voltage power supply sends an electrical discharge through a mixture of the two gases in a tube. This raises the energy level of the helium atoms, which then collide with neon atoms and raise them to excited states. The neon atom has a number of intermediate energy levels to which excited electrons can drop on their way back down to the ground state. The most widely used helium-neon laser radiates at 6,328 angstroms, in the visible part of the spectrum. Numerous other transitions in the neon atom can undergo stimulated emissions at specific frequencies in the infrared part of the spectrum.

Other types of lasers produce their laser light at other precise wavelengths, depending upon the energy transition levels in the atoms involved. Helium/xenon lasers emit at two wavelengths in the infrared, 3.5 microns and 2.026 microns. Cesium/helium lasers emit at one infrared wavelength, 7.18 microns. Argon lasers emit at a single wavelength of blue light, 4,880 angstroms. Neodymium/glass lasers emit at 1.06 microns, in the infrared.

Some lasers can be made to emit at multiples of the fundamental frequency, and some can be altered or tuned to emit over a small range of frequencies centered on the fundamental one. Dye lasers can be tuned continuously over the visible part of the spectrum. Optical parametric oscillators and semiconductor diode lasers can be tuned continuously in the infrared. But all these sources, especially the infrared devices, are relatively low in power and efficiency. The free-electron laser offers tunability across an enormous width of the electromagnetic spectrum at potentially higher powers and efficiencies. •

useful contribution to these areas," Brau acknowledges. "But even in the near term, free-electron lasers will be a useful source of coherent radiation for spectroscopy in wavelength regions not otherwise accessible, such as the far infrared." In fact, the first use of a low-power, free-electron laser for solid-state spectroscopy is expected to be under way before the end of 1982.

Patel, who chaired the National Research Council's subcommittee on free-electron lasers, offers several examples of ways in which special attributes of these new lasers are expected to be useful to him and his colleagues. Bell Labs does a lot of work in spectroscopy, the physics of interactions between matter and radiation. Of these interactions, many of interest occur at far-infrared wavelengths, and spectroscopists have long needed a source of intense, narrow, short-duration pulses of light of appropriate wavelengths to carry out the studies. But no conventional tunable lasers operate in the far-infrared region.

"What we do is limited by the intensity of the sources of radiation," says Patel. His group is able to use conventional sources of light, playing the light on the material under study sometimes for a full day to bring about a wanted interaction. "We would just wait," says Patel. "But that doesn't allow you to get any kind of time-dependent information." As is the case with a photographer seeking to capture a fast-moving event, if a scientist is trying to measure an event of microsecond duration, Patel explains, a pulse shorter than a microsecond is needed "or you can get no time resolution."

The free-electron concept

In producing a lasing effect by a narrow beam of electrons moving at relativistic speeds, the electron beam is directed past a series of magnets of alternating polarity. This magnet array is called a wiggler, or undulator; it causes the electrons in the beam to oscillate. As Charles A. Brau of the Los Alamos National Laboratory puts it, "They wiggle sideways."

When an electron is deflected from its path in this way, it gives up a small amount of its energy in the form of electromagnetic radiation, which may take the form of radio waves, X-rays, or light. So the electrons are radiating, and mostly in the direction of motion of the electron beam. But they are radiating incoherently; they are not organized in any systematic way; the radiation they emit is not in phase, or coherent, as is needed for laser action. (See "A laser primer," accompanying this article.)

One way to get the emission into phase is to get the electrons to organize into bunches. This can be done by superimposing a beam of light from a conventional laser on the electron beam. In the simplest terms, when everything is adjusted right, the application of the laser field to the electron motion causes the electrons to bunch and radiate coherently.

"They not only radiate coherently with respect to each other but also with respect to the applied laser field," says Brau. Their radiation adds coherently to the applied laser field and reinforces or strengthens it. This produces laser amplification.

The applied laser field that interacts with the electron beam can be from a separate, external, conventional laser, in which case the free-electron laser is an amplifier. Or the coherent radiation produced by the oscillation of the electrons can be returned through mirrors to the front of the system and passed back through to interact with the electron beam again and again, each time intensifying the light. In this case the laser output is also the input, and the free-electron laser is a self-contained system, an oscillator.

"Basically, what you've got in either case is a bunch of particles that want to radiate and want to radiate incoherently," says Brau. "In both, when you apply the electric field of a laser, they are made to radiate coherently." Another way to look at this is to view the free-electron laser as an electron accelerator working backwards. In the electric field of the laser, some of the electrons will be wiggling with the proper phase to be slowed down by the electrical force. The energy they lose is given to the laser field and amplifies it.

Unfortunately, you also have electrons entering the wiggler half an optical period later or earlier—out of phase. They will experience just the opposite electrical forces and will absorb light from the laser field rather than give up energy to it. Until this situation is changed, no net energy is gained by the laser field.

That is where the bunching comes in. Some of the electrons are accelerated by the interaction with the laser; others, half a wavelength away, are decelerated. As the faster ones catch up to the slower ones, they form bunches, coherently phased with respect to the light waves. With the electron beam formed into coherent bunches, net coherent energy is then extracted from the electrons in the form of coherent light.

It is estimated that with improvements in the technology, more than 10 percent of the electron energy might be extracted in this way and transferred to the light being emitted. •

A short pulse—say, 10 picoseconds, Patel explains—can cause the atoms in a material to absorb the energy and jump to a level higher than their ground state. Then another pulse, say 30 nanoseconds later, in effect sees the material at this higher energy state. A succession of such pulses can reveal how fast the atoms of the material under study come down from the excited state.

Another kind of question is how surface catalysts work. A catalyst is a layer of a substance that holds some other molecule in a particular orientation, so that it can enable a chemical reaction to take place. Consider the example of a monomolecular layer of carbon monoxide on a surface of nickel. Nickel is a good catalyst; the question is, in what orientation is the carbon monoxide molecule held by the nickel? Is the carbon atom down and the oxygen atom up or the reverse? Is the whole molecule lying flat on the nickel surface, or is there some in-between orientation? The answer is important because that will determine how carbon reacts with oxygen to make carbon monoxide.

"It all depends on which way the carbon is sitting there," says Patel, and a laser beam of the correct, tunable frequencies can help find out. Laser light of sufficient power and correct wavelength can measure absorption at the frequency of certain chemical bonds, in particular those between carbon-oxygen, oxygen-nitrogen, and carbon-nitrogen. "That tells you the geometry," he adds. For carbon monoxide, there are other ways of determining the geometry, Patel notes. But it is a good example. For many other molecules, the free-electron laser offers the only hope.

At lower frequencies

"We have had a crying need for a tunable high-powered laser in the far infrared field," Patel continues, especially in solid-state science. That is why Bell Labs is building a free-electron laser at its Murray Hill, New Jersey, site. The machine is expected to be operating by September 1982.

Some have called the Bell Labs project the world's first facility for free-electron laser users, but Patel says that description is a bit broad. It will not be open for use by the scientific community generally, but only by Bell Labs scientists. "Some collaborations are possible, but not on a demand basis," Patel says.

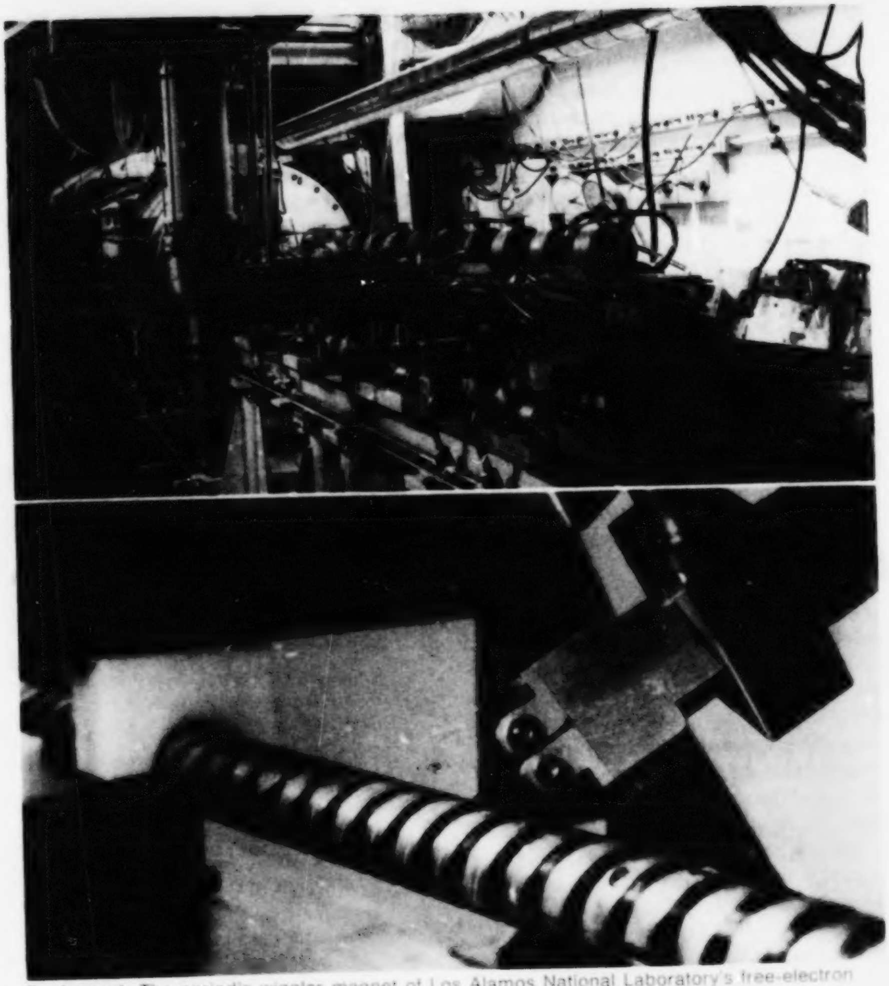
Nevertheless a number of physicists at Bell are expected to use it, and Patel says it will probably operate 24 hours a day. From Bell Labs' point of view, the need for a tunable laser source was most acute in the far-infrared region from about 10 microns to 1,000 microns. Since low-frequency free-electron lasers require far less energetic electron accelerators than do high-frequency ones, the relatively long wavelength tunable radiation desired by Patel and his colleague Earl D. Shaw requires relatively modest electron energies. As a result, a relatively economical accelerator called a microtron will be used for the Bell Labs facility. It is essentially a miniature cyclotron; Patel says it is small enough to fit on a desk.

Shaw and Patel say their free-electron laser will be capable of producing laser pulses as short as one to 100 picoseconds (10^{-12} to 10^{-10} seconds), with peak powers on the order of hundreds of kilowatts. It will significantly modify far-infrared spectroscopy. It will also make possible other experiments which cannot even be contemplated at the present time. It is, they agree, "a new frontier in science."

Despite all that, says Patel, "to us, the free-electron laser is just like any other laser source. We're using it because it can provide us with a capability we don't now have. We're not doing the physics of the machine itself. We're treating it as a tool no different from any other tools we use."

At the University of California at Santa Barbara, other physicists are greatly interested in the use of free-electron lasers in solid-state physics. They plan to put into operation in 1982 a free-electron-laser system that will make use of a Van de Graaff generator as the source of electrons. That laser "is unique among free-electron-laser devices," says Luis Elias who, with Vincent Jaccarino of Santa Barbara and William M. Yen of the University of Wisconsin, is developing it. "It is the only one that proposes the use of electrostatic accelerators of relative low energy and is designed for high powers as well as high efficiencies," he declares.

The California system will operate initially as a single stage free-electron laser in the wavelength range from 100 microns to one millimeter. This range is expected to be extended well into the vacuum-ultraviolet region, says Elias, using the technique he proposed in 1979 whereby two free-electron-interaction processes can take place using the same electron beam.



Equipment. The periodic wiggler magnet of Los Alamos National Laboratory's free-electron laser (top) and fabrication of the superconducting helical wiggler at Stanford University.

Charles Brau; John M[illegible]

The first stage consists of a conventional free-electron laser in which submillimeter coherent radiation is produced in a low-loss open resonator cavity. When the radiation field in this cavity reaches a high level of intensity, it acts as a short-period electromagnetic wiggler capable of forcing the same electron beam to produce much shorter-wavelength coherent radiation. Thus, says Elias, "the final tunability range of the Santa Barbara laser may cover a very broad range of wavelengths from one millimeter to 2,000 angstroms."

A six-million-electron-volt electrostatic accelerator will both accelerate the electron beam for the free-electron laser and decelerate the spent electrons to recover their energy. This is expected to increase overall efficiency and make high average power (as much as 12 kilowatts) possible.

Elias and the other scientists working on Santa Barbara's Free-Electron Laser Project expect to have the laser and all the spectroscopic and laboratory equipment assembled and operating by the end of 1982. They say they consider themselves to be at the threshold of a new era in submillimeter spectroscopy.

Another experiment being carried out, at the U.S. Naval Research Laboratory, is at still longer wavelengths. The goal here is to produce an efficient, tunable source of extremely high-power millimeter-wave radiation with possible application to plasma heating and millimeter-wave radar devices. A recently reported experiment there produced 35 megawatts of emission at approximately four-millimeters wavelength at a relatively high efficiency of 2.5 percent. The experiment used a relatively low-voltage (1.5 MeV) intense-current beam of electrons from an NRL accelerator.

The 35 megawatts is the highest power yet achieved by any device at these long wavelengths, according to Steven H. Gold of NRL. He says it indicates the potential of the free-electron laser to produce large amounts of tunable radiation in a difficult portion of the electromagnetic spectrum, one where neither conventional lasers nor the techniques using microwave tubes can produce high-power electromagnetic radiation efficiently.

At Brookhaven National Laboratory in New York, a free-electron laser is being built to operate at the opposite end of the electromagnetic spectrum, in the ultraviolet. Hope for the first free-electron ultraviolet laser rests with plans to insert a free-electron laser into the vacuum ultraviolet storage ring of Brookhaven's National Synchrotron Light Source. (See "Synchrotron Radiation II," *Mosaic*, Volume 9, Number 6.)

Storage rings

In a storage ring, the beam of accelerated electrons is not discarded after one use, as in a linear accelerator, but is maintained at high energies for a considerable length of time. With a free-electron laser inserted into a storage ring, the same electrons would be reinjected past the wiggler magnet time and time again, from one million to ten billion times. Moreover, the electron beam in a storage ring can be finely collimated. Such a capacity is considered necessary to make high-frequency laser radiation possible.

The storage-ring idea, says Claudio Pellegrini, who proposed the free-electron-laser project at Brookhaven, might be a way to produce laser radiation wavelengths as short as those in the far ultraviolet. One document prepared on the Brookhaven project even mentions the possibility of being able eventually to produce a laser at X-ray frequencies, beyond the ultraviolet.

"If all things work out fine, we could eventually have a free-electron laser working wherever there's a synchrotron radiation source," says Pellegrini. "It extends the capabilities of the synchrotron source."

John Madey's group at Stanford University also has plans to build its own storage ring especially for powering a free-electron laser. It is expected to be ready to go in 1985 or 1986. To test the theory for the storage-ring laser, the Stanford scientists joined a group of French physicists led by Yves Fauge at the LURE laboratory of the University of Paris. They used the free-electron laser in the ACO storage ring at Orsay, France. "The French provided the laser and the wiggler, and we provided our experience," says Madey.

The work, reported in 1981, resulted in the operation of the first storage-ring free-electron laser and the first visible-light free-electron laser amplifier.

Tapered wiggler

Scientists at Los Alamos National Laboratory and at several industrial research laboratories are also working to develop free-electron lasers of high power and high efficiency. To accomplish this, they have been exploring a new configuration of the wiggler magnet array. Their array is tapered from front to rear so that the periodicity of its magnetic field changes along the length of the array. In this way, the desired resonant interaction between electrons and light is maintained as the electrons passing between the wiggler magnets lose energy. This concept is most applicable to free-electron lasers designed for very high peak power, on the order of a gigawatt. The proportion of the electrons' energy extracted from the electron beam as laser radiation theoretically can be increased from a few tenths of a percent to several percent or more by this method.

Three experiments using this concept have been completed in the past year: at Los Alamos; at Mathematical Sciences Northwest, Inc., in Bellevue, Washington; and at

Lasing at x-ray frequencies

If X-ray lasers emerge from advances being made in free-electron lasers in the visible and ultraviolet parts of the electromagnetic spectrum, they would enhance fluorescent spectroscopy, X-ray lithography, and photochemistry. They would make possible X-ray holography for the working out of the three-dimensional structure of important biological molecules. They would show up in X-ray microscopes and in a range of communications devices.

Free-electron lasers at X-ray frequencies have not come out of important recent advances in conventional lasing techniques, however. Nor do the powerful, collimated X-ray beams produced by synchrotron radiation sources—along with lasing action tunable for the first time across the entire visible and ultraviolet spectrum—have the necessary interaction time to produce lasing in the X-ray region. And even were that available, the technology to produce the necessary high-quality, special semi-reflecting mirrors for lasers at X-ray frequency has been elusive goal.

John Madey of Stanford University's High Energy Physics Laboratory, who has made major contributions to free-electron laser development, believes, however, that both barriers can now be overcome and that a free-electron laser capable of operating at X-ray frequencies seems a reachable goal. He and his Stanford colleagues are considering one for possible insertion into the storage ring free-electron laser they are designing for operation principally in the visible and ultraviolet range.

Work at the Lawrence Livermore Laboratory on X-ray lasers recently came to light when the laboratory sought outside investigators interested in conducting X-ray holography experiments; details are unavailable. Madey's work at Stanford in no way overlaps Livermore's, however. "As far as this lab goes," Madey declares, "the interest we have is entirely scientific with respect to both design and application."

New developments in three areas have combined recently to lift Madey's optimism about a free-electron laser at X-ray frequencies. One is the advance in the theory and technology of the free-electron laser. Madey has worked out the mathematics and found that a basic requirement—the density of electron current—does not pose as serious a constraint to laser action at short wavelengths as was previously assumed. The second, accelerator technology, has advanced sufficiently to provide the necessary current density. The third, the technology for designing and creating mirrors to reflect a sizeable fraction (50 percent or more) of the X-rays striking them "is now well at hand," the Stanford scientist declares. The combination offers short- and long-range targets.

"The surprising thing is that the technology seems to be adequate [to achieve lasing action] to 100 angstroms and beyond." (The boundary between the far-ultraviolet and the soft-X-ray portion of the electromagnetic spectrum is a matter of somewhat fuzzy definition. From 10 angstroms down is clearly in the X-ray region. From 200 angstroms up, radiation looks like vacuum ultraviolet. Between 100 and 200 angstroms the definitions are vague. Madey is focusing on the region around 100 angstroms, and within a factor of two or three around

TRW, Inc., in Redondo Beach, California. All began with high-power, conventional, carbon dioxide lasers and used accelerated electrons of 20 to 25 million electron volts to amplify the laser radiation. All three experiments have reported encouraging results in tapping energy from a significant fraction of the electrons. Los Alamos and Mathematical Sciences have reported order-of-magnitude improvements in energy extraction over untapered wigglers, averaging as much as 3.5 percent in the Los Alamos experiments. This compares with 0.2 percent in the best untapered wiggler experiments done by Madey and his colleagues at Stanford. In addition, the Los Alamos results confirmed the theoretical predictions, thereby supporting projections of improved performance in the future.

The Los Alamos results were formally presented at a laser conference in April 1982, but an earlier announcement from the laboratory called the achievement "a major step in the development of a new laser that could revolutionize chemistry and has important implications for industry, energy conservation, and national defense."

Charles Brau says the improvement in efficiency from a fraction of a percent (even that figure is higher than for many conventional lasers) to several percent is an important advance. "That was the principle we demonstrated," he adds. Brau and his colleagues hope to improve the technology and, by recovering more electron energy, get the figure up to around 20 percent. They expect to have a complete free-electron laser system in operation at Los Alamos in 1984, and many investigators are encouraged. •

The National Science Foundation contributes to the support of the work discussed in this article through its Quantum Electronics, Waves, and Beams Program.

that figure. He regards the 10-angstrom level as "very challenging.")

Large current

On reason for Madey's optimism became evident only when he sat down earlier this year and worked out the current density requirements for a free-electron laser to function at short wavelengths. Short-wavelength free-electron lasers need a fairly large current from an electron accelerator, and the current needs to be restricted to a fairly small cross-sectional area and what is called phase-space volume.

This combined current density requirement was expected to go up rapidly as the desired lasing frequency increased toward the X-ray end of the spectrum. Actually, says Madey, the rate of increase turns out to be not so great; the required current density goes up in proportion to the reciprocal of the wavelength raised to the half power. To get a factor of a hundred difference in wavelength requires a factor of ten difference in current. "The dependence on wavelength is really pretty flat," says Madey.

When he next plotted the current density required to achieve lasing action at short wavelengths, around the X-ray end of the spectrum, Madey found that, as expected, linear accelerators do not provide the necessary current density. But storage rings do. The synchrotron radiation produced in storage rings permits fractional damping that forces the electron beam to a smaller phase space. This greater current density offered by storage rings of one billion electron-volts energy is sufficient to allow lasing action down to at least 100 angstroms. Two storage rings with that ability already exist: the Stanford Linear Accelerator's damping ring and Brookhaven National Laboratory's new vacuum ultraviolet ring. Madey expects his proposed ring to do far better.

Madey's figures are based on a 100-percent gain for each time the electrons pass through the laser. (Gain in a laser is the ratio of power out to power in.) Of the X rays produced in the laser, half may be lost through the mirrors; a gain of 400 percent per pass may be necessary. But that doesn't rattle the investigators. "If you need 400 percent gain per pass, you have to move up in your current density requirements, but it still seems likely that you can reach well beyond 100 angstroms," Madey says.

Mirrors

While the technology and theory of free-electron lasers and accelerators have been moving forward, the same can also be said of mirror-making abilities. Mirrors placed at the two ends of laser cavities are made of thin alternating layers of materials. These stacks of layers are designed to provide constructive interference of the electromagnetic waves as they reflect from different depths, thereby providing enhanced reflection. Mirrors for an X-ray laser require finding pairs of materials that will reflect X rays perpendicularly and then synthesizing the layers to thickness tolerances as fine as half an angstrom. Scientists at IBM began to design normal-incidence mirrors for X-ray microscopy in 1974. Work underway since 1976 at Stanford and other laboratories for a wide variety of applications has now reached the level where mirrors for X-ray lasers appear feasible. "I was surprised to learn that the mirror technology is as advanced as the accelerator technology," says Madey.

Troy W. Barbee Jr. of Stanford's materials science and engineering department is optimistic about his laboratory's ability to make the mirrors for the laser Madey envisions. "The technology is such that one can expect to make normal-incidence mirrors for wavelengths down to 30 angstroms," says Barbee. He says new calculations by Alan E. Rosenbluth of the University of Rochester identify a number of promising materials pairs; alternating layers of rhodium and silicon look especially promising. "At normal incidence for 124-angstrom wavelengths, rhodium-silicon should be able to achieve larger than 80-percent reflectivity," says Stanford's Barbee.

Madey and Barbee marvel at the way the accelerator technology, free-electron laser technology and theory, and mirror technology all have suddenly coverged to the point where now it appears that a free-electron laser can be designed to produce laser radiation at the soft X-ray wavelengths. "It is a remarkable coming together of technologies," says Madey.

Although a formal X-ray laser program has not been established at Stanford, "there is a strong enthusiam for it here," Madey says. "The people are in place." Madey envisions having Stanford's new free-electron-laser storage rings, now being designed, modified to accommodate insertion of a lengthy straight section. The new storage rings and the free-electron laser would, Madey says, be ready in about five years, with an X-ray laser "perhaps a year after that." •

UNDER BABEL'S TOWER

by Lee Edson

How children acquire language is a riddle for developmental linguists and the subject of debate among them. The acquisition process may be universal.

At no time in the history of psychology has so much attention been paid to baby talk. Teams of psycholinguists record the early words of Georgian youngsters in the Caucasus, eavesdrop on the chatter of Bantu-speaking infants in the Congo, preserve the prattle of toddlers learning Quechua in the Andes, and contend with the Polynesian language in Samoa and Tagalog in the Phillipines. Even deaf children are being videotaped to observe their gestures as they learn to communicate.

The object of all this multilingual babel in sound and symbol is to determine not only how children acquire their native tongues, but to find out if they do it in the same way and to uncover what the answers reveal about the human brain.

Child language as a window on the mind has been of interest to philosophers and scientists in one way or another since ancient times. But the first important modern psychological theory of language acquisition, at least in the United States, did not appear until after World War II, when psychological thinking was dominated by behaviorism. B. F. Skinner, the leading American exponent of behaviorism, explained language as the product of stimulus-response interactions explicable by what is considered standard learning theory.

Most of the behaviorists' experimental work in language that could be carried out in the early 1950s started with nonsense syllables or random materials that were far from the structure of language. Subjects, both children and adults, were presented with lists of arbitrary pairings of words and syllables. They were asked to learn as many as they could in a given period of time then tested on their recall. Their rates of learning and what they learned were measured.

Out of such work the behaviorists concluded that the children were learning language by associating words with objects, by imitating their elders, and by a Pavlovian force called conditioning that is at the base of behaviorism. None of the learning theorists could explain the real mystery of language: why two-year-olds acquire language so rapidly and why four-year-olds develop the ability to combine old words into a seemingly infinite string of new statements.

A universal skill? Some linguists argue that children acquire language through a universal process regardless of their native tongues. Clockwise from lower left: children in India, Japan, the United States, and West Cameroon, and (over) Russian children in Moscow (top) and German and Turkish children in a Frankfurt classroom.

Jagdish Agarwal, David Ryan, Bruce E. Remer, Evan G. Schneider—Uniphoto, by permission.

A new idea

The time was ripe for a new idea, and it arrived in the late 1950s in the form of linguist Noam Chomsky's hypothesized language-acquisition device; this and other related theories have come to be called the Chomskyan revolution. The language-acquisition device, according to Chomsky, is an inborn apparatus of the mind—a kind of cerebral faculty—that contains a full complement of linguistic tools and blueprints. It enables a child to learn the language of his parents regardless of their nationality or the language they use.

When confronted with verbal input, the cerebral faculty recognizes its character and formulates rules about its structure. It then generates the grammatical representations needed for comprehending the input and establishing the grammar required to achieve communication in the language being acquired. In biological terms, the faculty might be compared to the immune system, which goes into action to generate from an almost infinite repertoire of possible antibodies the one that will recognize a specific antigen.

Chomsky thus created a new view of language acquisition as a system of inherited rules for constructing a grammar. The grammar enables the individual to express what Chomsky called the deep structure of a sentence. This represents the meaningfulness that lies beneath the expression. One familiar example of this is to recognize that some sentences have the same meaning, or deep structure, even though surface structures, or manifest forms and order of the words, differ. For instance: "Dog bites man," or "Man is bitten by dog".

Chomsky's transformational theory of grammar inspired a generation of psychologists to test the empirical validity of his ideas, and methodologies began to flourish in the 1960s. Children were at the focus of most of the research. The tape recorder was the technological advance that helped make the research manageable.

Out of the melange, three major groups took center stage. At Harvard University, Roger Brown and his colleagues carried out an important longitudinal study of several youngsters. Susan Ervin-Tripp and Wick Miller undertook a similar study at the University of California at Berkeley; Martin Braine of New York University directed the third study, using five children, one being his own son. Whether the body of utterances, or corpus, as linguists call it, was representative of the population at large is open to dispute; it was made up of utterances of the progeny of Ph.D's. Moreover, the general emphasis was on English.

Nonetheless, developmental linguistics, as the field came to be called, was on its way. Since then the field has exploded, aided by such 20th century tools as the computer, the video recorder, and the jet airplane—a not inconsiderable piece of advanced technology that made many remote peoples readily accessible to researchers.

Universal patterns

English began to give way to other target languages. At Berkeley, for instance, Dan I. Slobin heads a cross-linguistic study of language acquisition that has enabled him to look at the data on no fewer than 56 languages. They include such exotic tongues as Bhojpuri (Indo-Iranian), Garo (Sino-Tibetan), and Ga (a language of Ghana). He has studied 15 of them in detail.

At the University of Illinois, Charles Osgood and his coworkers use a network of linguistic sources to provide data from languages around the world. At Stanford University, Charles Ferguson and Joseph Greenberg—who have compiled the world's most extensive collection of language universals—have been able to offer this data base as a guide to their colleagues.

As an indication of growth in the new field, there is an international association that studies child languages and holds triennial conferences. In several European countries, universities have established child-language centers. And in the United States, two major institutions—Boston and Stanford Universities—hold annual, widely attended conferences on language acquisition. A major periodical, *The Journal of Child Language*, is being published. The field's findings continue to be essentially descriptive. Nonetheless, scientists see in this newly created

Wide World Photos, by permission.

body of information the beginnings of a theory of language acquisition. It is one that applies universally; it is as valid for Bantu as it is for English.

The pattern that has emerged seems to follow a series of clearly marked stages. Some time around a child's first birthday, the first linguistic utterances occur. They consist of one or two syllables, most often of the form called consonant-vowel: da-da, ba-ba, or the like. How much meaning do these words carry? During the 1960s, Eve Clark of Stanford, Katherine Nelson of Yale, Lois Bloom of Columbia University, and Jerome Bruner of the New School for Social Research, then of Harvard, tried to find the answer.

Bruner, for instance, found that children discover words at transition points in their activity. When they see new things, they attach words to them. But the words are more than labels. Children signify location, use negatives, and issue single-word comments such as the verb "open" when seeing a jar of jam. They also generalize. Any liquid may be referred to as milk, for instance. These meaningful sentence-in-embryo words are known to psychologists as holophrastic speech, and they often seem to be so insightful as to suggest a personal comprehension beyond the sense of the words.

By age two

In the second stage of speech development—at about 20 months or so—children begin more frequently to put words together to express simple notions and actions. A structure in line with reality starts to emerge. Children have not only broken up the stream of sound that was their repertoire in the first two years; they have also begun to perceive that two words together express a different meaning than do single words.

This phenomenon has attracted a good deal of psycholinguistic research and has led to some interesting if short-lived theories. In the 1960s, for instance, Martin Braine introduced the concept of the pivot word, a member of a two-word combination that is constant, the other word being variable. The child might say "more milk" or "more talk". Indeed, the child might produce such utterances as "more book" (Don't stop reading) and "more car" (Don't stop driving).

The concept of the pivot word excited linguists because it seemed to open the possibility of a universal child grammar. But the idea didn't stand up against a barrage of subsequent findings.

Lois Bloom, for instance, pointed out convincingly from her own observations that more than Braine's combinations of constants and variables was involved in early childhood talk. Some children, she said, did not make the combinations Braine had found in his own subjects, and other children made other combinations. Bloom argued that the Braine hypothesis was faulty because it was too simple, that children actually can create grammar as complex as an adult's.

Braine, while disagreeing with the extent of Bloom's proposed grammar, subsequently modified his hypothesis. He came up with the idea that children acquire simple rules for expressing meaning. He called his hypothesis a limited-scope formula. What it says is that if an English-speaking two-year-old wants to say someone is doing something, the child puts the actor in the first position followed by the word that reflects the action: "Daddy eat," for example. If the child wants to convey the location of an object, the object is named and then the location, using a construction like "book table". There might also be a formula for possession (mommy purse), for class (mommy lady), and for size (big car).

"Evidently the children proceed from rules specifying semantic roles for each word," contends Braine. As to word choice and order used, Bloom and Braine agree that children may vary in the rules they acquire first, depending on how they perceive the world.

The international child

More recently, Slobin's cross-language studies have shown that two-word utterances are pretty much the same in all languages, and they represent the same kind of idea development. For instance, negative expressions like "no wet" in English have such counterparts as *nicht blasen* (not blow) in German, *vody net* (water no) in Russian, *ei suzi* (not wolf) in Finnish, and *le ai* (not eat) in Samoan. All show similar advances in the child's linguistic capacity rather than in the grasp of new concepts. Says Slobin: "They represent more of a universal look than they do a Finnish or Samoan look."

How does the international child move on to discover and build grammar? Consider the expression that suggests location, such as "book table" or "pot stove," a kind of combination that appears in all languages. In English, the child shows an understanding of which object is doing the acting or is the focus of the expression and which object is acted or located upon. This is shown by means of the word-order regularities of speech: The child almost never says "table book," for instance, or "stove pot".

Position

In other languages, Slobin and his colleagues have found that the principle is the same, except that identification often occurs by means of a grammatical marker or inflection, sometimes on the subject and sometimes on the object. Slobin, for instance, has compared the Indo-European languages such as French, Spanish, or Italian, with two other great language families, the Altaic languages, which include Turkish, Korean, and Japanese, and the Uralic languages such as Finnish and Hungarian. He has noted an important difference.

In the Altaic or Uralic groups, certain actions are communicated by suffixes or post-positions (words following the noun modified), rather than by prefixes or pre-positions (words ahead of the noun) as in the Indo-European languages. In Hungarian, for example, noun suffixes regularly indicate the position of an object and its movement: *hajó* means boat, *hajóban* means located in the boat, *hajóbol* refers to movement out of the boat, and *hajótal* is used to express the act of moving away from the boat. Similarly in Turkish, an end-of-noun inflection is used to express the English equivalent of "pot stove on" or "wine bottle in".

According to Slobin's study, suffixes and inflections work for all children in these languages as word order does in English. Not only that, but the position of the marker is important to the process of acquisition; markers placed after the noun (suffixes and post-positions) are easier for children to acquire than markers placed before the noun (prefixes and pre-positions).

This is true in less familiar languages, too. In the case of SiSwati, a language spoken in Swaziland (and using both prefixes and suffixes), E. C. L. Kunene reported in a dissertation at the University of California at Los Angeles that children learned suffixes earlier than they learned prefixes. It seems that children find ends of words and utterances more memorable than beginnings and middles.

Other language components are also learned at different rates, depending on the structure of the language acquired. For example, take the three-word Turkish sentence *O beni güldürdü*, which can be translated as "She made me laugh". But as Slobin points out, the Turkish language uses bits of meaning that are attached together as single words. Thus in literal translation into English, the Turkish sentence *O beni güldürdü* actually reads: "She I to laugh make did," in which the last four bits form a single verb. This type of construction is learned at an earlier age than the English equivalent in which meaning remains spread out over several words.

Once children have isolated and acquired elementary grammatical forms, they must go through the steps needed to determine what Slobin calls "the scope of the grammatical form's application." That is, the depth of meaning is expressed either by the position of a word or by a prefix or suffix. Here again, in all languages the child at two begins with some limited-scope formula, which must gradually be modified to conform to the norms of adult language.

For instance, consider a simple sentence made up of a subject (actor), verb (action), and object (receiver of the action), in that order. According to Slobin, an English-speaking child is alerted to who does what to whom by word order. In Russian, an inflection is added to the object to identify it as the direct recipient of the action. The word "man," or *chelovek*, appears as *cheloveka* when "man" is the direct object of a verb, as in "Touch the man". Russian two-year-olds easily acquire this suffix.

Why a difference?

However, according to the late A. N. Gvozdev, a leading Soviet authority on child language in the period after World War II, a Russian child applies the ending only when the object is physically affected by the action expressed in the verb, such as occurs with verbs like "hit," "carry," or "embrace". When *chelovek* is the object of a verb that is less active, like "see" or "hear," the children do not apply the suffix. Adults always use *cheloveka* with such verbs.

The reason for the distinction is not known. Slobin notes that the children appear to have discovered that a grammatical element encodes a tangible notion, in this case the notion of acting physically on a physical object to change its state or position. This is something that occurs frequently in the world of children, and their grammar coincides to a certain extent with their perceptions. The grammar, as Slobin sees it, is not yet generalized and abstract.

Similar acquisitions are found in many other languages, even in those in which coding is entirely different from that of English or Russian. The Kaluli tongue of Papua New Guinea is a case in point. Kaluli is one of the so-called ergative languages, a group that includes Basque, Hindi, and American Indian tongues. In ergative languages, the object of the verb does not receive special grammatical marking but the agent does.

In these languages, the marking is attached to the word denoting the subject, but this happens only in sentences in which the subject is doing something that has an effect on something else. For instance, in such sentences as "Mother feeds the dog," "Father spanked the child," and "The pig eats the mud," the words "mother," "father," and "pig" receive the markers. If someone acts without effect on an object (for example, "Mother is sleeping"), the ergative case suffix is not used; the subject remains unmarked just as the object is unmarked in transitive sentences.

Amazingly, children have no trouble learning this unusual relationship at the age of two. And they never make the error of extending the marker of a transitive verb's subject to the subjects of such intransitive verbs as "run," "jump," and "fall," no matter how active their behavior might be.

Then, at four

Later changes also appear to show across language groups. As youngsters reach the age of four, for example, another major step in language acquisition takes place. In all cultures children go through a period during which they strive to analyze and improve the clarity of their native language. In doing so, they make mistakes that have been the subject of several cross-linguistic studies. Annette Karmiloff-Smith, a student of child psychologist Jean Piaget, working in Geneva with French-speaking children, describes an experiment in which a child was asked to move all his toys into a gas station and relate what he was doing. That child said *Mes voitures doivent aller au garage* (My cars must go into the garage), whereas a child two years older described the action as *Toutes les miennes de voitures doivent aller au garage* (All of mine of the cars must go into the garage).

Apparently in the second phase the child wants to make sure each element of meaning—totality (*toutes*), possession (*miennes*), and plurality (*de voitures*)—is emphasized separately, even if the grammar is wrong in adult French. The grammatical alteration of the younger child's expression nonetheless results from the effort to imitate adult speech rather than from the child's own internalized grammar. As the child grows older and begins to form his own grammar, however, phases will be gone through in which patterns diverge again from adult forms.

Groping with language, the scientists believe, is part of an overall strategy children develop to extract meaning from the word arrangements and inflections they encounter. To test this hypothesis, Slobin, Tom Bever of Columbia University, and their colleagues conducted a series of experiments with two- to four-year-olds who spoke English, Italian, Turkish, or Serbo-Croatian.

Each child was shown a pair of toy animals and told to act out a certain task, like mimicking a dog biting a squirrel. The instructions were given in various word sequences. Sometimes the subject was placed first in the sentence, and sometimes it was second or last. Each child was graded on the ability to recognize the agent, the performer of the action, and the object of the action, no matter how the set of words might have been structured.

Slobin and Bever found that the Italian- and English-speaking children respond generally to noun-verb-noun sentences, a form prevalent in those languages. However, there are differences in response at different ages. At two, the youngsters do not respond systematically to word order and do not develop a strategy for dealing with it. Later the children go through a state in which they sometimes get the meaning wrong. They tend, for instance, to call the first noun in the sentence the actor, even in sentences where this is not the case, as in "The horse was kissed by the cow".

By age four this behavior is beginning to change. The children then judge the words in a sentence presented to them in terms of meaning, not just order; as Bever's earlier work shows, they may transform passives into actives. They are beginning to understand content and to form grammatical strategies that will lead to adult language.

By contrast, Turkish youngsters pay little or no attention to word order but respond primarily to inflections, to word-form changes that provide the linguistic clues to the subject or object of the sentence. Nonetheless they do it at the same speed as the English and Italian children do.

In terms of accuracy in performing the test, the Turkish children were superior to the English and Italian children. Among other economies that Turkish offers is that Turkish children do not have to transform passive constructions in their minds. On the other hand, the Serbo-Croatian children encountered the greatest trouble among the groups tested. Their language allows more word-order variation than does English or Italian. It is not as highly inflected as Turkish, however, making the task of Serbo-Croatian children more complicated.

Sound, syntax, rhythm

As children mature, they also acquire a strong grip on the sounds within their languages and there are patterns of adoption of sounds as there are of grammar. Studies in phonology show that children acquire sounds in sequences and rates comparable to those at which they acquire syntax and semantic understanding. Although the process is still a mystery, current research does show that all children start out with similar sounds. They learn along the way to develop the differences that make the French child speak French and the English child speak English.

How do children actually master the sound system of the language? Apparently, according to recent evidence, they know facts about pronunciation that are not displayed in their speech. Braine says youngsters have what he terms a phonotactic filter, an innate clumsiness that makes it hard for them to reproduce certain sounds or combinations of sounds that they hear. Thus an 18-month-old may say "pik" (pig) or "gog" (dog) but will object to such inaccuracies in someone else's speech.

The mechanism of this filter has been studied in several different ways and on different levels by various investigators. At Stanford, for instance, Charles Ferguson, a pioneer in the study of sound acquisition, recently did comparative phonological studies among English-, Spanish-, and Cantonese-speaking children. He focused on one phonological feature: the contrast between voiceless sounds such as *p* and *t* which use only breath and tongue or lip, and voiced sounds such as *b* and *d*, which require vibration of the vocal cords along with breath.

The three languages vary in the manner in which the voicing is used; Ferguson found that although the first stages of acquiring

---

Edson is the author of six books and many articles on science.

the voicing are common to children regardless of the language acquired, the speed of development depends on the language.

In attempting to delve deeper into the development of pronunciation, George Allen of Purdue has looked at speech rhythms in several languages and highlighted the similarities and differences in the way rhythms are developed. One thing he found quickly—as have linguists studying other features—is that part of learning consists of forgetting. Children display at one stage certain vocal accenting abilities that disappear as they get older, because the vocalizations are not part of the language being acquired.

French children, for instance, can detect a regular type of rhythmic stress difference on a syllable at the age of four, but it represents a pattern that does not occur in French, and they cannot detect it a year later. "The ability is not really gone, as evidenced by work with Swedish and German children," says Allen. "The process is a little like losing baby teeth. The early sound patterns disappear and are replaced by others—the ones that remain."

It may be that children remain free to express such universal patterns because their parents have other concerns. Roger Brown once expressed this notion succinctly by stating that "children learn the rules of the language despite the parents."

He found that a parent will seldom correct a child's grammar, but will correct the content of a child's statement. If parents were solely responsible for the learning of language by their children, the children might be left to grow up, as one researcher puts it, "as truthful illiterates."

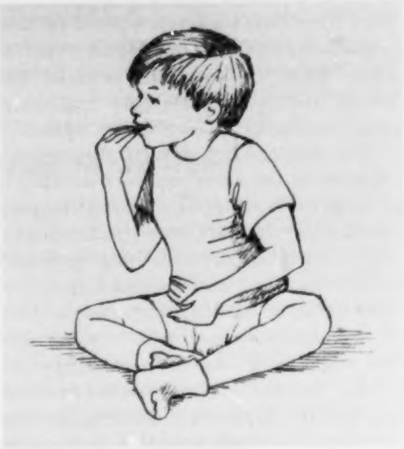
The cross-linguistic studies have shown this to be true of several cultures. In societies as diverse as the Luo of Kenya and the Mayans of Yucatan, children are generally not corrected when they make errors in grammatical form. Yet studies show that these languages are acquired at the same rate and in the same sequence as is English. Moreover, errors of a type that adults seldom make do appear in the talk of many children and sometimes persist nearly to adulthood. This would be unlikely if children were learning by imitation.

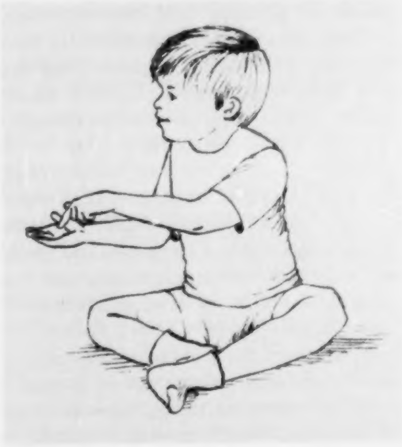
Social interactionists

Even if the imitation of parents is not the primary process of language acquisition, a number of psychologists suggest that some form of social interaction between parent and child plays an important role. The philosophical father of social interaction theory, Lev Vygotsky of the Soviet Union, argued back in the 1930s that language evolves out of children's adjustments to socially appropriate ways of speaking. "Language," he said, "is a merger between outer speech the child hears and inner speech he thinks with."

Developing sign language. Linguist Susan Goldin-Meadow (top left) has observed that deaf children will develop a gestural language of their own, even with no one to imitate. Such gestures can include, for example, bringing fingers to the mouth to mean "eat" or "food" or pointing to an upraised palm to mean "give". The signs in sequence mean "give me the food".

Patricia Evans; Noel Yovovich/U. Chicago





Since that statement was made, social interactionists have followed various paths, differing by and large on where they place their emphasis. Some researchers, notably Elizabeth Bates of the University of California at San Diego, have looked at the intent of the speaker in communication and how intent molds language. Others, such as British linguist M. A. K. Halliday, have examined the functions of communication in assisting children to accomplish their social goals through language.

Still others, including Jerome Bruner of the New School and John Dore of the City University of New York, have tried to find a single clarifying concept that ties language and communication together. They, along with Daniel Stern of Cornell University Medical College, have done this by focusing on what they term the entire speech act, which includes intent and function along with emotions and other communicative building blocks of language acquisition.

Finally, exponents of what is called motherese, including Catherine Snow of Harvard and Jean Berko Gleason of Boston University, have singled out the important role of the mother in the acquisition of language. According to these investigators, the mother is the one who expands the child's utterances into appropriate language by encouragement and reinforcement. "Language acquisition," one scientist says, "requires joint problem solving by mother and infant."

Even the deaf

Special evidence of the innateness of the language capacity has appeared in studies by Susan Goldin-Meadow of the University of Chicago and others working with deaf children. Goldin-Meadow's vehicle is a study already eight years long of deaf children whose hearing losses have prevented them from acquiring oral language naturally and whose hearing parents have not yet exposed them to conventional sign language. Despite this, Goldin-Meadow has observed, the urge to communicate is so great in the children that they develop a gestural language of their own at an early age even when they have nobody to imitate. "They appear to invent signs to specify objects and actions," she says, "and concatenate these signs into longer strings to convey semantic relations."

The deaf children develop their gesture systems in the same sequence as hearing youngsters develop their spoken language systems. At the start, they point to objects and characterize actions—bringing fingers to the mouth in a drinking motion to represent the act of drinking, for example. This, says Goldin-Meadow, is the equivalent of the single-word stage of the hearing child. At age two, the deaf children put two simple gestures together—the drinking sign plus a gesture toward the drinker—a development paralleling a hearing child's two-word stage.

These gestures are not at all like the signs of conventional sign languages. They tend to be more transparently related to the essential features of the objects they denote—a twisting motion to represent a jar for instance—than are the signs of American Sign Language. Even in what are called oralist schools, those in which attempts are made to teach deaf children lip reading and speaking and where signing is discouraged or suppressed, the children develop signs which they use behind the teacher's back.

By age four, the deaf children develop an order and complexity of gestures that again parallels many of the word-acquisition accomplishments of hearing children. To determine whether such gestures are imitations of observed gestures, Goldin-Meadow compared mothers' gestures to those of their children. The mothers' gestures were either not structured or were structured differently from those of their children; in either case, they could hardly provide a model from which the children could induce the structure of their own gesture systems. Goldin-Meadow also studied the mothers' responses to the children's gestures to see if they could have been shaping the structural properties of those gestures, but she found no evidence for such a shaping process.

Similar findings have been obtained with deaf students who learned American Sign Language. Elissa Newport of the University of Illinois points out that second-generation deaf—the deaf children of deaf parents, who number about 10 percent of the deaf population—develop sign language far richer in morphology than that of their parents. Indeed, fresh aspects of American Sign Language are observed to be invented by every generation of the hearing impaired.

In the future the study of language acquisition may move in directions that do not depend entirely on observations of children. At the University of California at Irvine, psychologist Kenneth Wexler and linguist Peter Culicover assume that all natural languages are learnable and ask: What properties must a language have to make it that way? To answer this question they have devised the basics of a computer model involving different ways of arranging learning systems. With appropriate instructions, a computer might be able to unravel the grammar of a language from a program containing its surface structures.

None of these findings, of course, overshadows the fact that to a preponderance of psycholinguists, inheritability remains the discipline's core question. Chomsky proceeds from the belief that the child's brain inherits a good deal of information about language, information that is called upon as the ability to talk develops. The cross-linguistic studies with their new findings on universals do not prove or disprove this.

Children don't know in advance the form their thoughts will produce, but they can process linguistic data. The overriding question of whether this capacity springs from general cognitive abilities or is a faculty of its own like vision will remain a point of contention until the evidence on one side or the other becomes overwhelming. •

The National Science Foundation contributes to the support of the research discussed in this article through its Linguistics Program.

The Regulation of Gene Expression

by William Check

The triggers that turn genes on or off appear to involve movement, relative position, chemical modification, and nearby proteins.

The development of all progeny of a single fertilized egg into the varied cells and tissues of an adult animal is a baffling process. Why do red blood cells make hemoglobin, specialized cells in the pancreas secrete insulin, and nerve cells synthesize neurotransmitters such as acetylcholine and dopamine?

All human cells contain genes for hemoglobin, insulin, and acetylcholine. Why do only a minority of cells make these chemicals? The difference should be explicable in terms of expression of the respective genes.

Although mechanisms in the triggering of a gene's expression are being uncovered, the heart of the question remains elusive. In the last few years, however, the tendency of several lines of research to converge—or at least to relate their results to one another—is increasing both the level of excitement among some of the scientists involved and the rate at which models are coalescing.

One line of evidence suggests that the difference between a gene that is being expressed and one that is not is that the silent gene is chemically modified. Added to it in strategic places is a methyl atomic group—it is methylated. There is also evidence that active genes are extremely sensitive to an enzyme—DNase I—that cuts DNA. More recently scientists have discovered the presence of two proteins that attach preferentially to genes that are being transcribed; they may enhance attachment of other factors necessary for transcription or may somehow stabilize an active conformation of genes.

On a more complex level, one crucial distinction between the active and inactive states of one particular set of genes—studied in frog eggs—is the presence on the active genes of large aggregates of proteins, called transcription complexes, that are unable to coalesce on inactive genes.

Finally, a group of yeast geneticists investigating the expression in a yeast cell of the gene for one sex or the other has found that the ultimate determinant is a second, entirely separate set of genes on a second chromosome. The function of this second set of genes is not clear, but it involves the mobility of gene segments, an event found frequently in various organisms in the last few years.

One encouraging aspect of the recent flowering of knowledge in gene expression and control is that the scientists involved are actively thinking about the ways all these different lines of evidence might fit together. The biochemists are asking whether the absence of methylation, the presence of sensitivity to DNA-cleaving enzymes, and the binding of proteins might all be different facets of the same process. It is even conceivable that they might be sequential steps, all essential to gene expression. The geneticists wonder whether the regulatory genes in yeast, for example, act by altering methylation or by affecting the sensitivity of DNA to a cleaving enzyme.

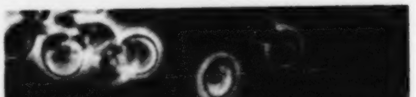
If this sudden wealth of genetic, cytologic, and biochemical information about gene regulation can be assembled into a unified theory of differentiation, that theory could explain several current mysteries in considerable detail. These would include the switching on of some genes early in the embryonic stage while others switch later, during development. It might also show how somatic cells extend to their offspring selectivity for those genes that need to remain in the expressed state—so that, for example, the progeny of skin cells will always have as their progeny other skin cells.

Yeast: the best known

To investigate this subcellular prescience, scientists need to know more detail about what genetic elements are important in transmitting this selectivity. To do this, many turn to yeast, whose genetics is among the best studied in higher organisms. And at the Cold Spring Harbor Laboratory on Long Island, one group of geneticists is painstakingly unraveling the complex system by which yeast regulates its sex.

Yeast does not seem very sexy to humans. In fact, the two yeast genders are not even identified as people label theirs: male and female. Yeast genders are just *a* and *alpha*.

But the system is as well developed as that of higher animals. The two genes known to control expression of these two genders are called mating-type alleles. (Alleles are alternate forms of the same gene that result



Yeast cell division. Common baker's yeast exists as single round cells about three microns across. The dividing zygotic cells are produced as a result of mating between cells of opposite sexes, or mating types, called a and *alpha*. The system producing a cells or *alpha* cells is a consequence of gene regulation.

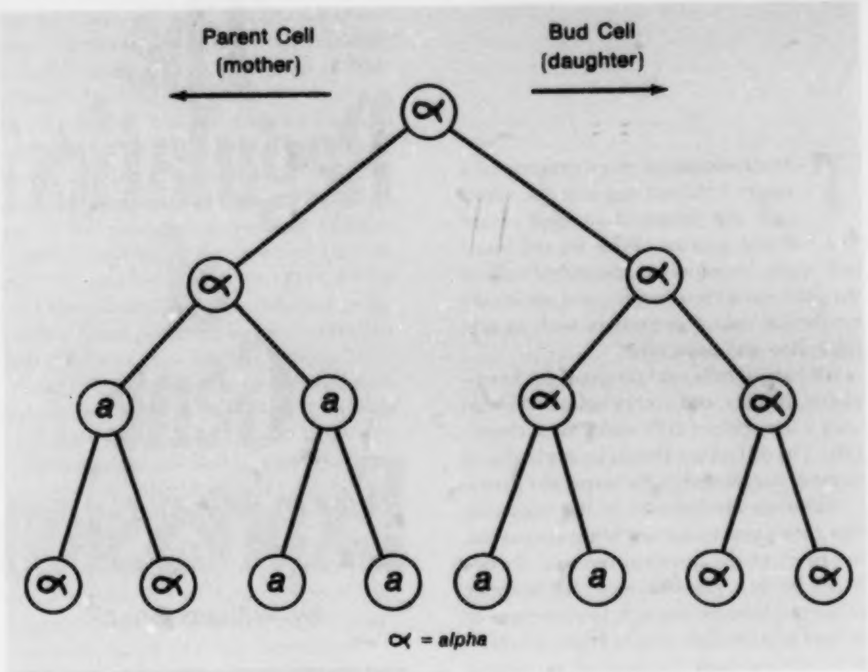
Amar Klar, James Hicks, Jeffrey Strathern in different properties in the organism.) Each yeast mating-type allele dictates whether a cell will be an *a* or an *alpha*. The allele also regulates expression of several genes on other chromosomes, genes that express such secondary sex characteristics as differing cell-surface properties, different sensitivity to mutagens, and, most important, the synthesis of one of two 13-amino-acid pheromones.

One pheromone is produced by *a* and the other by *alpha*, and they truly act as promoters of mating in yeast. Each type of pheromone inhibits growth of a cell of the other type at a particular point in the cell cycle. This synchronizes the cells' processes so that they can mate. Mating gives rise to zygotes, each diploidal—carrying two sets of genes. These cells go through two cell divisions each to yield four spores, two *a*'s and two *alphas*, each a haploid, having one set of genes. The spores germinate and each gives rise to a colony of yeasts.

It is when an *a* or an *alpha* cell is allowed to grow in the absence of the other cell type that the yeast shows its sexual diversity. In the first division, the spore gives rise to two cells of the same type, say one *alpha* dividing into two *alphas*. But the two cells that arise from this division are not equal. Under microscopic observation it becomes clear that one gives rise to the other; one is a mother cell and one a daughter.

In its next division, the mother cell undergoes a dramatic change. It spontaneously gives rise to two cells of the opposite mating type: two *a*'s. One of these cells is the mother in the pair, and it can keep switching mating type at each subsequent division. The daughter cells from each division go through the same process as the spore cells: one round in which they yield two cells of the same mating type and a second round in which they switch.

As unwieldy as this system might seem, says Amar Klar, one of the Cold Spring Harbor researchers, it represents a "predictable, stable, directed system of gene switching." Klar and his colleagues, Jeff Strathern and Jim Hicks, see the yeast mating-type system as a simple differentiation model. They are exploring the ways in which these genes are switched on and off in successive generations—processes that involve, for instance, gene motility.



A switching pattern. Yeast mating types switch in successive yeast cell divisions: Mother cells switch mating type at each subsequent division; daughters, on the other hand, give rise to two cells of their own mating type.

After Amar Klar, *et al.*

Cassette models

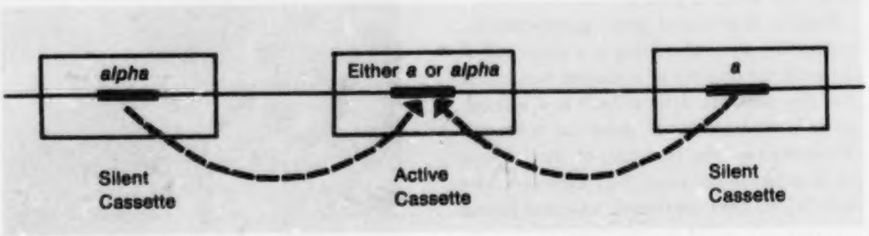
Experiments done at the University of Oregon by Hicks, Strathern, and Ira Herskowitz and by Klar while at the University of California at Berkeley have disclosed the role of the motility of DNA segments in the gender determination and reproduction patterns of baker's yeast (*Saccharomyces cerevisiae*). The scientists found that a yeast cell that started out with a defective mating-type gene somehow came up with a good version during the mating-type switch.

Further analysis revealed that yeast cells have three copies of the mating-type gene. One is expressed, two are silent. One silent gene is *a*, the other *alpha*. When a budding cell becomes a mother cell, the gene in the expressed, or active, position is replaced by a copy of the mating-type gene from one of the silent positions. In standard laboratory strains, 90 percent of the time the replacement gene is of the opposite mating type, leading to switching of gender.

The Cold Spring Harbor scientists think of the active gene as a normal cassette that can be recorded on and played back. The silent genes are special cassettes that cannot be played or recorded over. But they provide the information that is the basis for replacing the mating-type gene in the active position. (See "Gene Segments on the Move," *Mosaic*, Volume 12, Number 1.)

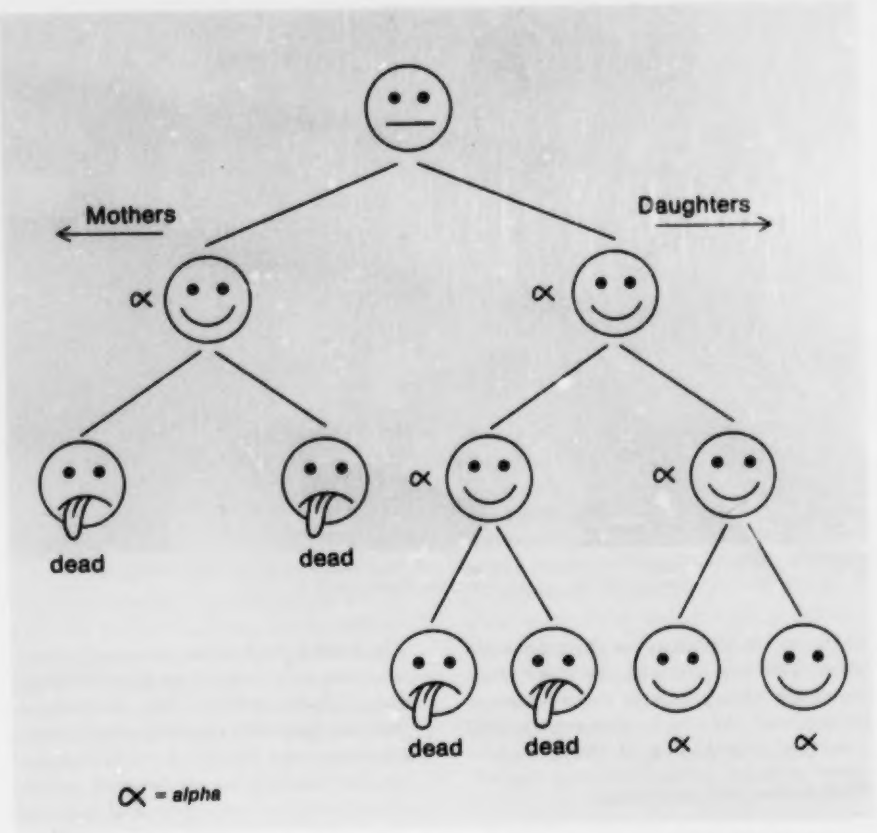
More recent work has shown that the analogy to an audio cassette can be carried further. On the back of an audio cassette there is a tab. When the tab is broken off, the recording ability of the cassette is eliminated. Klar, Strathern, and Hicks have found such a positive control system in yeast.

Both they and Herskowitz have found a set of four genes that regulate the expression of the silent cassettes. Herskowitz calls them silent-information regulators, while



Cassette model. The active, or expressing, cassette gene is flanked on the chromosome by silent cassettes, one containing an *a* mating-type gene and the other an *alpha*. The switch of mating type most often involves replacement of the mating-type gene in the active cassette with a copy of the opposite mating-type gene from its silent cassette.

After Amar Klar, *et al.*



Manipulated for study. Genetic surgery produces dead yeast cells in a predictable fashion, making yeast a model for the study of gene regulation.

After Amar Klar, et al.

Klar calls them mating-type regulators. Whichever, when one of these four genes is mutated the silent cassettes can express their mating-type genes. They can also accept a mating-type gene from either the active position or the other silent position. As a result, one can think of these regulator/repressor genes as being responsible for breaking off the tab on the silent cassettes so that they are limited to transferring information only to the expressed cassette. When one of the regulating genes is mutated, the tab is not broken off and the silent cassettes become active.

These insights enable scientists to phrase questions about regulation more pointedly: How do the regulatory/repressor genes keep the silent loci silent? How is a mating-type gene transferred from a silent site to the expressed site? What activates the expressing site and causes the organism to replace *a* and *alpha* mating-type genes at every other cell division as it does? The first two questions are yielding at least partial answers; the third is still a mystery.

Some answers

Postdoctoral fellow Judy Abraham in the Cold Spring Harbor group has obtained a partial answer to the first question. She placed a silent *a* cassette on a plasmid, mutated it, and inserted it into a diploid yeast cell in which both mating-type genes were *alpha*. (Since the two mating-type genes were the same gender, the cell could not express its diploid potential.) Some of the silent *a* cassettes on the plasmid were mutated in such a way that the *a* gene was expressed. These cells became true *a/alpha* diploids. They could form spores and they lost their initial ability to mate with *a* cells.

On analysis of the changes in these initially silent cassettes, Abraham consistently found loss of DNA in one of two specific regions around the mating-type genes. Apparently these regions of DNA, called E and I, must be intact for the four regulator/repressor genes to keep the silent genes silent.

It is not yet known how the regulator/repressor genes act on the E and I regions to prevent expression of the silent cassettes. It may be through an effect on DNA conformation that the mating-type genes in the silent cassettes are made inaccessible to the enzymes necessary for expression. This concept of different conformational states of the chromatin is often invoked by scientists working on gene regulation, and many intriguing and reasonable models have been built around it.

Kim Nasmyth, working as a postdoctoral fellow in the group, has found that the DNA in silent cassettes on a plasmid is more tightly coiled when the regulator/repressor genes are active. This type of evidence has led Jeremy Thorner of the University of California at Berkeley to suggest that the product of the regulator/repressor genes may be a topoisomerase, an enzyme that facilitates DNA twisting. (See "DNA Topology: Knots No Sailor Ever Knew," *Mosaic*, Volume 12, Number 1, and "The Structures That Make DNA Work," *Mosaic*, Volume 11, Number 1.)

Strathern has found a different gene that can help answer the second question: How are mating-type genes in the silent cassettes transferred to the expressing site? The gene he has located (called homothallic, or HO) is responsible for directing the switching of mating-type alleles. Yeast strains in which this gene is altered (heterothallic genes) retain the same mating type at all times, as if they lacked silent cassettes altogether.

During transposition of the mating-type gene from the silent to the active site, Strathern has determined, there was a double-stranded cut in the DNA of the active site, right at the end of the mating-type gene. This cut was directed by the HO gene, and heterothallic strains did not make the cut. The piece of DNA at the end of the mating-type gene (and the cut made there) is crucial, Strathern found. If this small section of DNA is removed, no switching occurs.

Klar further demonstrated the biological importance of the double-strand cut, using strains in which the silent cassettes had been excised by genetic surgery. Mother cells in these strains produce two dead cells rather than two live cells of the opposite mating type. The cut is made in the DNA of the active cassette, but it cannot be repaired in the absence of the silent cassettes and the DNA is destroyed. No cut is made in daughter cells, so they give rise to two live progeny of the same mating type.

One further finding uncovered by Strathern was an interaction between the HO gene and the regulator/repressor genes. Ordinarily the HO product cuts only in the active cassette, even though all three contain the vulnerable stretch of DNA. But when regulator/repressor genes are defec-

---

Check writes principally on biology.

Seeking answers. James Hicks, Amar Klar, and Jeffrey Strathern (left to right) are trying to unravel the complex system by which yeast regulates its sex.

Joan James

tive, cuts are also made in the silent cassettes. This explains why the silent ones can switch when regulator/repressor genes are mutated: The locus that experiences a cut acts as a recipient of information.

Methylases and nucleases

While geneticists have been working out these complex interactions in the yeast mating-type system, biochemists have been investigating gene regulation in higher organisms. They are also looking for chemical differences between expressed and inactive genes. One of the first such differences found: More of the cytosine nucleotides in an inactive gene have a methyl group attached. (The four nucleotides—cytosine, guanine, adenine, and thymine—are the basic building blocks of DNA and the informational elements of the genetic code.) The methyl molecule, a single four-atom group, is found attached throughout the DNA.

One of the earliest teams to discover that placement of these methyl components is not random was headed by Gary Felsenfeld and James McGhee at the National Institutes of Health. They thought that differences in methylation would be found only at special sites. Using enzymes that cut DNA specifically at these targeted sites, McGhee looked at the DNA of chicken embryos, which do not make the protein globin, and at adult red blood cells, which actively synthesize it. McGhee found a high level of methylation in the globin genes of the embryo. In contrast, a large fraction of the same targeted sites were unmethylated in the globin-producing adult red blood cells.

According to McGhee, this was "a clear-cut example of lack of methylation being related to gene activity." But, he cautions, "we only showed a correlation, not a cause-and-effect relationship." A further correlation has been uncovered between the undermethylation of active genes and sensitivity to a DNA-splitting enzyme, DNase I. These sensitive sites do not occur at random throughout the gene; some researchers think they are clustered at the end where transcription begins. In addition, so-called hypersensitive sites in this early region of an active gene can be cleaved by several other types of DNA-splitting enzymes.

At the Fred Hutchinson Cancer Research Center in Seattle, Hal Weintraub and Mark Groudine have expanded these findings. They showed that in at least one case there appeared to be a more direct relationship between undermethylation and gene activation. They used chicken cells containing an unexpressed virus called ev-1, which is in one of the chicken chromosomes. In contrast to the DNA of the active virus ev-3, the DNA of ev-1 is heavily methylated. Groudine and Weintraub, in collaboration with Bob Eisenman, put chicken cells with ev-1 in them into a medium containing a chemical that interferes with methylation of the DNA. When they did so, ev-1 was expressed and a virus was produced.

This team also uncovered what appears to be a correlation between methylation and gene expression in chicken sperm. Most of the DNA in these sperm cells is heavily methylated, says Groudine. This includes permanently inactive genes as well as those that are tissue specific and will be expressed in the adult chicken. But, as in other systems, the continuously expressed genes are undermethylated near their beginning.

Groudine cautions that undermethylation "is not a cut-and-dried explanation" for gene activity. In several cases undermethylation does not correlate with globin gene activity. As an example he cites chicken embryo fibroblasts. When these cells are infected with Rous sarcoma virus, their globin genes switch on. But so far the scientists have not been able to detect any change in the state of methylation around the globin DNA. "Unfortunately," Groudine observes, "there's a generalization gaining credence that methylation is the main factor in gene expression. I would say that undermethylation is not sufficient for gene expression and may or may not be necessary."

On the other hand, Groudine points to the presence of hypersensitive sites as a "necessary but not sufficient" condition for gene transcription. In the chicken embryo fibroblasts, for instance, activation of globin genes by the Rous virus is accompanied by induction of hypersensitive sites. Such sites also arise during induction of ev-1 virus.

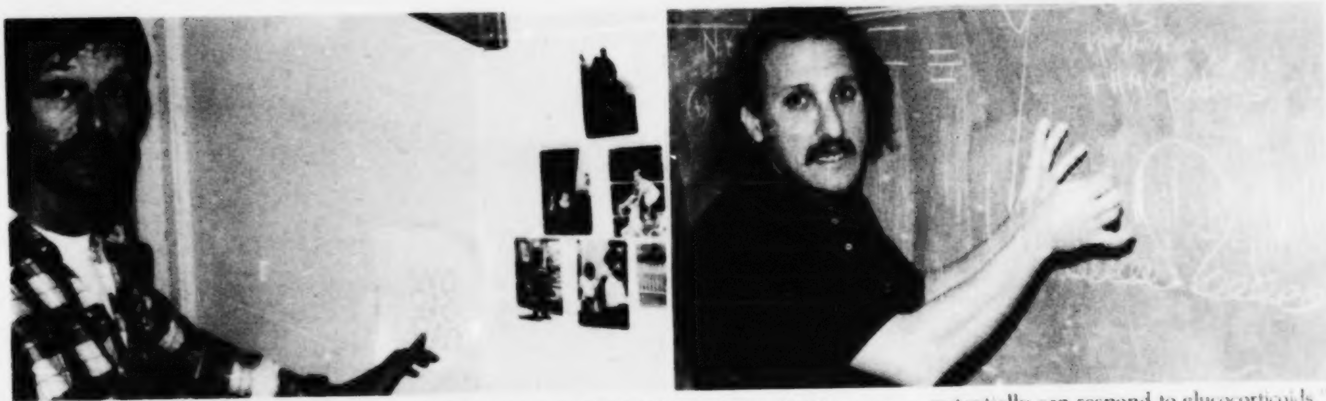
Collagen genes show a similar pattern. The work of Benoit de Crombrugghe and his colleagues at the National Cancer Institute has showed that, in almost all tissues, the early region of collagen genes is undermethylated. Yet only some tissues express the gene. In situations where the collagen gene is activated, the difference seems to be the presence of a hypersensitive site at the beginning of the gene.

"There are probably many parameters that define gene activity," Groudine suggests. "If a gene is already undermethylated, then introduction of a hypersensitive site may be the decisive factor. But it may be a rare gene for which there is a single difference between the inactive and active state."

A pair of HMGs

A more recently discovered factor as promising as hypersensitive sites is a pair of proteins prosaically named HMG 14 and HMG 17, for the technique by which they are isolated. They are a part of a group of proteins that move rapidly in an electric field, hence HMG for high-mobility-group proteins.

Initially this entire group was perceived as a mysterious mix. Then in 1979 Stuart Weisbrod in Weintraub's laboratory found that two of the proteins, numbers 14 and 17, were reliably associated with actively transcribing globin genes. Their presence also accompanied gene activity in systems in which methylation was not a differentiating



Looking to biochemistry. James McGhee of the National Institutes of Health (left) and Mark Groudine of the Fred Hutchinson Cancer Research Center in Seattle are studying the chemical differences between expressed and inactive genes.

William Check

factor, such as the turn-on of globin genes in chick embryo fibroblasts by Rous sarcoma virus infection. Why these two proteins are involved and not others, however, is not known.

Now working in John Gurdon's laboratory at the Medical Research Council in England, Weisbrod has used the two HMG proteins as tracers to isolate nucleosomes, the protein-DNA assemblages in which genes are arranged in the cell nucleus. The process selectively pulls out nucleosomes attached to active genes, implying that the HMG-nucleosome interaction may be an important factor in the initiation or the maintenance of transcription.

The work of Groudine and Weintraub especially seems to pull together many of the inferences coming out of other laboratories. "We obviously think that changes in chromatin structure are very important for gene regulation," Groudine says. "One example where this seems to be true is the mating-type locus in yeast, where regulatory elements may produce a long-range effect by inducing form changes."

He and Weintraub have obtained indirect evidence that regulation of the chicken globin genes may similarly be effected through changes in chromatin structure. They looked at DNase sensitivity of the genes for ovalbumin, embryonic globin, and adult globin in the three types of cells: red blood cell precursors called stem cells, embryonic red cells, and adult red cells.

In all three types of red blood cells, ovalbumin genes are inactive and resistant to DNase. Weintraub and Groudine interpret this as meaning that ovalbumin genes are in a condensed or tightly coiled chromatin region. In the precursor cell, both types of globin gene are inactive, but they have an intermediate nuclease sensitivity; this leads to the speculation that they are in a relaxed chromatin domain, a preactivation state.

In the embryonic red cell, embryonic globin genes are expressed and these genes are now hypersensitive. Groudine and Weintraub interpret this as being due to addition of activating factors, such as Weisbrod's high-mobility group proteins. The adult globin genes are still inactive and remain of intermediate sensitivity. This indicates that, though they are in the relaxed chromatin domain, they do not have the HMGs and are still methylated.

In adult red blood cells, adult globin genes are expressed and are hypersensitive, but the embryonic globin genes have become insensitive to DNase. This may mean that they have been reeled back into the condensed chromatin region.

Developmental models

Several scientists are testing the genetic and biochemical information from the yeast and globin studies in other developmental systems. For example, Richard Palmiter of the University of Washington and Ralph Brinster of the University of Pennsylvania have generated the first system in which a gene introduced into a mouse embryo is correctly regulated in the adult animal. They showed experimentally that in this system a gene can be introduced into an egg and developmental control can be subsequently exerted. As a result, the system offers the possibility of studying the gene's condition—methylation, DNase sensitivity, level of expression—at stages of development.

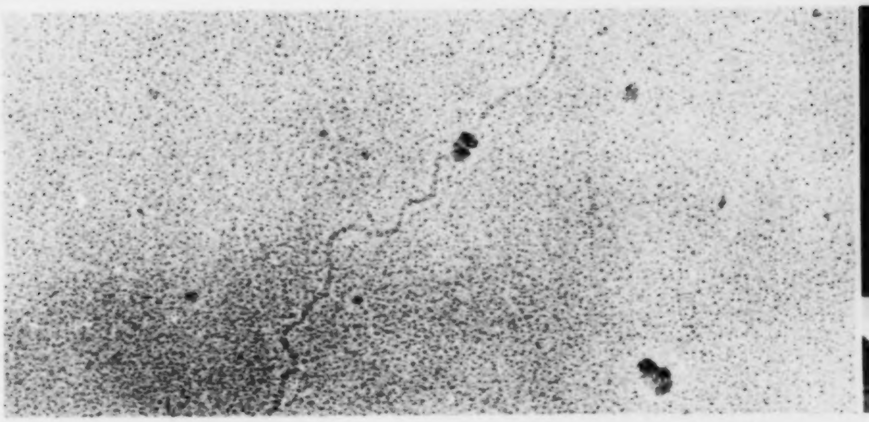
An equally versatile system is being explored in Keith Yamamoto's laboratory at the University of California at San Francisco. He uses hormone-inducible expression of the mouse mammary-tumor virus as a microcosm of the whole developmental process. "There is a set of genes in every cell that potentially can respond to glucocorticoids," Yamamoto explains. "But in any cell type, only a subset does respond." He is defining this hormone responsiveness at the gene level.

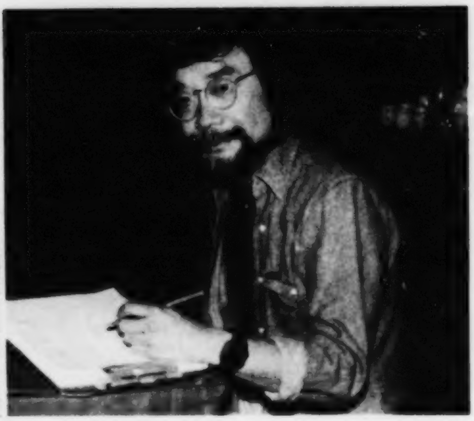
Fortunately for the scientists, a readily manipulable segment of DNA that encodes mouse mammary-tumor virus responds to glucocorticoids. It remains silent in the chromosome until induced by a hormone. The hormone binds to a receptor protein within the cell and the hormone-receptor complex is translocated to the nucleus, where it binds to the chromosome. How does the complex interact with the chromosome? And what changes does this binding cause to bring about virus transcription?

Yamamoto and his colleagues found that the complex binds selectively to several sites on the viral genome itself. To explore how these areas function, the researchers excised them and put them on plasmids next to a mammalian gene not ordinarily subject to hormone regulation. The receptor binding sites made the gene hormone responsive, Yamamoto's group found.

Yamamoto thinks that the binding sites probably are not acting as promoters, since they are separated on the virus genome and most of them are far from the promoter site. "We keep thinking about changes in chromatin structure," he says in words similar to Groudine's. "Perhaps receptor binding makes the chromatin more accessible to RNA polymerase." His group is now testing whether the receptor acts like high-mobility group proteins to induce DNase enzyme hypersensitivity within or near hormone-responsive genes.

Yamamoto and his coworkers had previously found that the hormone inducibility is highly correlated with the DNase sensitivity of the chromatin flanking the virus in the genome. Further testing showed that the virus genes seem simply to adopt the chromatin structure of the region of the chromosome into which the virus inserts. Thus the virus is expressible if it integrates





Defining hormone response. Keith Yamamoto, of the University of California at San Francisco, works with mouse tumor viruses that contain sites to which hormone receptors bind. The lumps on the DNA strand are hormone-receptor complexes.

William Check; Keith Yamamoto

into a chromosomal area that is hypersensitive, an indication that it is in a permissive state for expression. This makes the viral system an excellent probe for exploring the relationship of chromatin structure and regulation of gene expression.

Where in the flanking region does DNase sensitivity change? Using viral DNA as a probe, Yamamoto's group has isolated a border between a DNase-sensitive and a DNase-resistant segment of the liver-cell genome. The border apparently represents the transition between repressed chromatin and chromatin that can be transcribed. He suggests that there must be genes that do something to maintain the chromatin in these specific structures. If a gene responsible for maintaining the repressed structure at a given site became defective, the chromatin at that site—and the viral genes inserted into it—would become free to express their messages. Mutant cell lines like these are now in the process of being isolated.

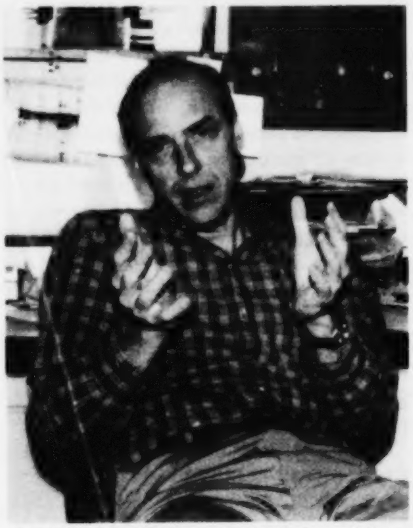
In addition, Yamamoto and his coworkers have cloned border regions and are inserting them at random sites in the chromosomes of various cell types. Determining under what situations the border sequences are recognized and used should give some clues to the way the cell decides which genes are to be expressed.

What might be the function of the genes that keep the border zones repressed? The subject is speculative, but Yamamoto raises the analogy to the yeast and the repressor/regulator genes. The scientists working with yeast are getting a handle on genes that specify the state of a DNA region, he points out. There may be similar factors that act on mouse mammary-tumor virus genes in mammalian cells.

Auxiliary genes

A special developmental problem and model—regulation of the genes that make a small RNA species (5s RNA) in frog eggs—is under study by Donald Brown at the Carnegie Institution laboratories in Baltimore. This RNA is a necessary part of ribosomes, the centers of protein synthesis in cells.

During its one-celled state, the frog egg goes through a period when it makes ribosomes at a rate equivalent to that of thousands of ordinary cells. Yet it has only one nucleus. To make enough 5s RNA to go around, the egg has developed a set of 80,000 auxiliary genes for 5s RNA that are used only at this time. It also has a smaller 16,000-member set of 5s RNA genes, but they are turned on only when the frog reaches embryonic and adult stages.



Studying frog eggs. Donald Brown of the Carnegie Institution is exploring centers of protein synthesis for gene regulation.

William Check

One important mechanism for regulating these genes is a protein that is necessary for their transcription. The factor is plentiful in the egg. It becomes greatly diluted, however, when the egg goes through a series of rapid divisions to form an embryo without further synthesis. The protein appears to bind more tightly to the smaller set of genes than it does to the auxiliary set.

How this factor works may have implications for how genes can be stably turned on or off. That makes it, in a nutshell, a model for stable inheritance of a cell's phenotype. Donald Brown has shown that the promoter factor binds not at the initiation site of the 5s RNA gene, but in the middle. At this site the factor gathers around it a series of other proteins that form a stable complex. "The RNA polymerase loves these complexes," says Brown. It attaches to the internal complex and then reaches back to the initiation site to transcribe the gene.

But if the complexes are formed without the protein factor, they are inactive. In the presence of normal chromosomal proteins called histones, an inactive complex becomes permanently fixed in the inactive state. Since replication does not change the status of complexes on a gene, Brown proposes, these findings can provide a model for stable differentiation states.

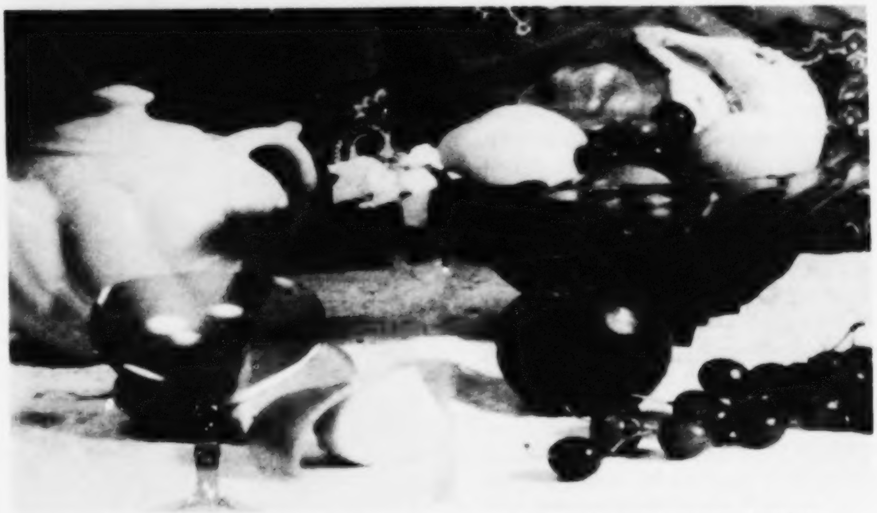
Still, Brown does not push for the validity of this or any other mechanism for regulating gene expression. "All methods of regulation are still up in the air," he asserts. "The fundamental truth about modern biology is that it is principally an empirical science. Most insights are gained from experiment, few from theory." Given the variety of experimental systems now being used to explore gene regulation, surely many more insights are imminent. •

The National Science Foundation contributes to the research discussed in this article through its Genetic Biology Program.

A Palette for Computers

by Allen L. Hammond

Computer-generated imagery is reaching a level of sophistication that promises to challenge the most discerning eye. The power of available computers may be a limiting factor.



Added piece. A champagne glass incorporated by computer into a digitized photograph is so well integrated it refracts the image of the pitcher behind it.

Donald P. Greenberg, Doug Kay/Cornell

Computer graphics is outgrowing experimentation and coming into commercial use at a rapid rate. It comes in a variety of forms: Computer-aided design is already a multi-million dollar business. Most military and airline pilots train with simulators based on computer-generated images. There is an upsurge of systems to produce charts and graphic displays for business managers. And computers are beginning to play a role in art and in entertainment including animation sequences in such films as "Star Trek II" and "Tron."

Underlying these applications are increasingly specialized hardware and a huge assortment of programming techniques. Among these is a sophisticated set of mathematical procedures, or algorithms, especially devised for creating and manipulating images in a computer. The goal of research in this area is still-more-realistic computer-generated images.

Compared to the production of familiar wire-frame and other such images characteristic of conventional computer graphics (see "Computer-aided design," accompanying this article), the generation of realistic images requires an enormous amount of computation. Television-type display screens most commonly used, for example, are typically divided into a square grid with 512 sub-areas on a side. Each is an individual picture element, or pixel. The process of creating a realistic image must eventually come down to calculating for each pixel the intensity of each of television's three primary colors (typically red, green, and blue), the mixture of which produces tones and shades. The total can be nearly a million bytes of data for each picture. Yet calculating intensity values for pixels is only the last of a very lengthy series of manipulations needed to create and display a realistic picture; still finer grids are coming into use.

The process of image development begins with the creation by a programmer of a three-dimensional mathematical model of the objects to be displayed—say a group of buildings defined by the surfaces that produce their shape. These surfaces would be displayed initially as edges, however. (Though they are far more complex, they look like wire-frame images, and that term is used to describe them at this stage.)

The programmer can then rotate the image to find the best point for viewing it. Portions of the image that are backward-facing, behind another object, or otherwise hidden must be found by mathematical proceedures and eliminated.

Then the image must be shaded by computing intensities of light at each point on the surfaces, taking into account the location of light sources, the color and distribution of reflected light, and even such features as surface roughness and the material of which the surface is made. A brick building, for example, looks more grainy than one made of wood, even if both are the same color. Finally the image must be converted from its mathematical representation within the computer to discrete pixel units ready for display on a screen.

It sounds like a recipe for computer overload, and it is. "What you really want to do in computer graphics," says Charles Csuri of Ohio State University, "seems to require more computing power or speed than we have by a factor of ten to a hundred, and it's always been that way." And as computers have improved, so have the complexity and realism of the pictures attempted by computer graphics scientists.

The demand has kept ahead of the supply. The result is a field in flux. It is only in the last few years that faster computers and inexpensive, fast computer memories have made it practical even to attempt to display complex images or subtle color effects. But general-purpose computers, even large ones, are still too slow for many applications, and the cost of many computer graphics techniques is still too high for widespread use. The field's response currently is the reintegration into computer graphics, for some time predominantly a software field, of the design and development of custom hardware.

Efforts are under way to devise parallel processing schemes that would speed calculations, to encode part or all of the display algorithms in custom circuitry, and to exploit the inherent parallel processing capabilities of very large scale integrated silicon chips. This new hardware emphasis is in turn forcing researchers to rethink and recast software in forms compatible with the new hardware implementation, or to try radical, innovative approaches.

Persistent problems

One of the curiosities of computer graphics research is that its intellectual foundations are widely acknowledged to have been largely at one university: the University of Utah. The work centered there from roughly the mid-1960s to the mid-1970s. David Evans, then head of Utah's computer science department, attracted Ivan Sutherland from MIT, to teach and do research in computer graphics, a field that at the time had large ambitions but little to show. Sutherland had an enormous influence on subsequent research, and Utah produced a string of graduate students that includes many of the present leaders of the field.

Many of the generic display algorithms now in use were first developed at Utah. They include several methods to find and eliminate unseen portions of an image (hidden surface algorithms), methods to shade and display optical reflections, and methods to deal with color, surface texture, and the phenomenon known as aliasing, which causes the jagged or stairstep edges often seen in computer graphic images. The field has broadened in the past decade to include many research centers and an even wider range of applications. But many of the early themes have remained.

One continuing theme arises from the need to synthesize more complex images quickly and even automatically. How are images best represented within the computer? What mathematical forms should be used to model buildings, or automobile bodies, or mountains? Such problems as these have preoccupied computer scientists from the beginning and still generate lively arguments.

The earliest and still most common approach is to represent objects as a set of polygons, one for each planar or near-planar segment of each surface. This modeling technique works well for rectilinear objects such as buildings, and even many curved surfaces can be approximated by polygons if they are small enough and enough of them are used. Polygons have a further advantage: They can be represented simply in a computer. This is done by giving the three-dimensional spatial coordinates of each vertex along with a list of the vertices that are connected. This simplicity in the data structure translates into speed in processing.

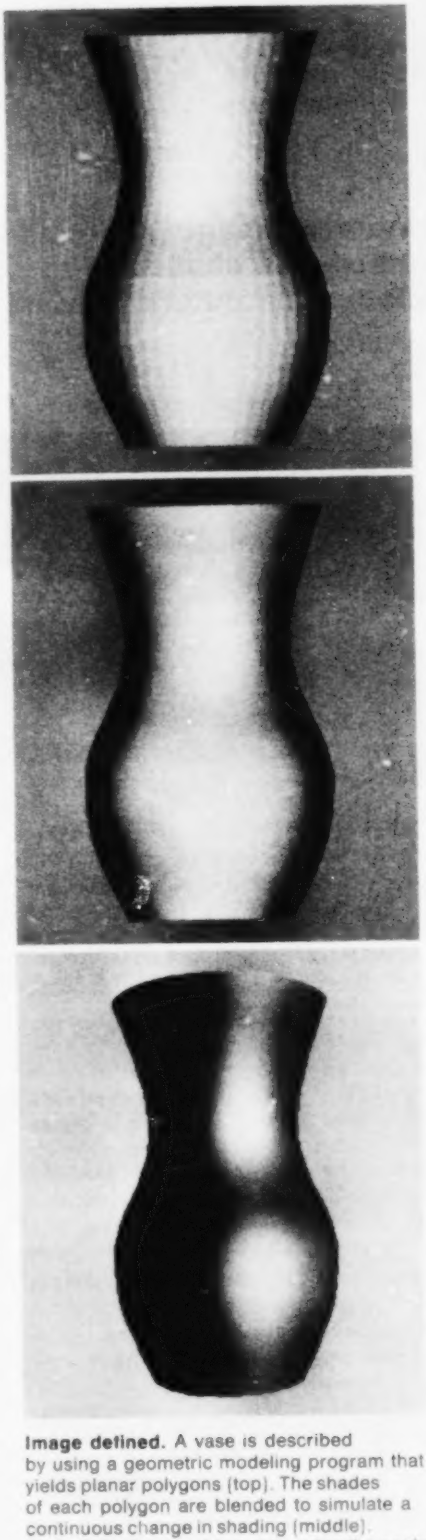


Image defined. A vase is described by using a geometric modeling program that yields planar polygons (top). The shades of each polygon are blended to simulate a continuous change in shading (middle). Diffuse shading is added, tailored to the angle between the source of the light being reflected and the observer.

Donald P. Greenberg/Cornell

But polygons do not readily represent objects with complex curves—automobile bodies, for instance. Nor does it usually make sense to create thousands of polygons if an object can more easily be described by other programmable geometric shapes. So conic curves (second-order, or quadratic-surface, modeling) are also used, along with solids of rotation built out of them, such as spheres and cylinders. More general curved shapes are modeled (third-order, or cubic-surface, modeling) with more complex surfaces that can be assembled in space to form patches. These are continuous surfaced segments that can curve in two directions at once as saddles or fenders do. Several research groups are developing software to exploit these higher-order models. Being included are efficient algorithms to subdivide or find intersections between patches and to manipulate them for display.

Essential features

To capture the essence of many natural objects—mountains are recognized, for example, by their ruggedness and clouds by their puffy or wispy texture—yet another modeling technique is needed. Benoit Mandelbrat of IBM's Watson Research Laboratories recognized some years ago that certain stochastic, or probabilistic, properties of many types of natural phenomena can be expressed by a mathematical property known as fractional dimensionality; he called these expressions fractals, or fractal sets. Loren Carpenter of Lucasfilm Limited in San Anselmo, California, Don Fussell of the University of Texas at Austin, and Alain Fournier at the University of Toronto came up with ways to apply this to graphics.

In effect, the algorithms map the underlying polygon or curved patch onto itself in a random but controlled way to create the jagged edge or rough surface desired. This works extremely well for many types of terrain. Fussell and Fournier are looking at other kinds of functions and combinations of stochastic and more ordinary deterministic modeling to represent clouds, trees, and ocean waves.

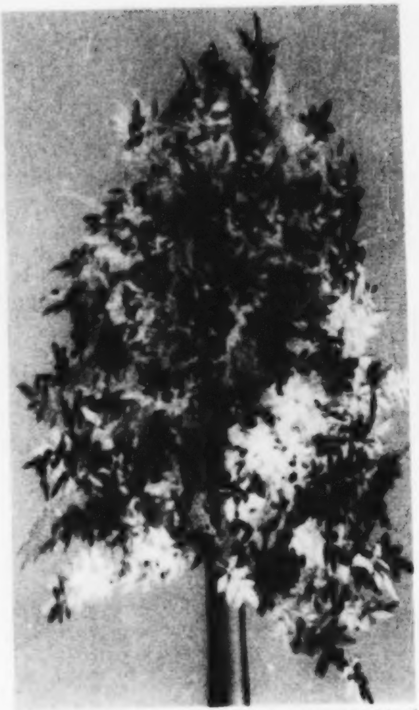
Another powerful technique for those applications that permit it is procedural modeling. In this, the geometry of the object is not explicitly defined by the programmer but is generated by the computer. A recent example developed by Robert Marshall at Ohio State University takes a single leaf as drawn by a programmer, scales it up or down in size, and rotates it with the aid of a random-number generator. In this way Marshall produces randomly varying leaves to form a tree. The model tree is then similarly scaled, rotated, and replicated to construct a forest. The resulting picture of a forest can contain up to a thousand trees and 40 million edges, and it can require up to 11 hours to compute. When done, however, it attains a degree of realism that far surpasses the output of simpler, more stylized models.

Once a model exists, of course, it often serves as the basis for many images such as those encountered in a walk around a building or a flight past a landing strip. A second go-round can be accomplished by simply reapplying the display algorithms to different viewing points of the same three-dimensional model. How much computation it takes depends on several factors: the complexity of the image as measured by the number of polygons it contains, the degree of sophistication and realism attempted in the display algorithms, and the kind of hardware used to perform the computations. "You usually need at least one polygon per pixel for a picture to look real to the human eye," says Lucasfilm's Loren Carpenter, "and that's where we're headed."

At present, with conventional minicomputers, the display of a simple image containing several hundred polygons can take a few seconds; more complex images on the same equipment can take hours. Simulators in which algorithms are implemented by custom-made, high-speed circuitry can display several thousand polygons in a thirtieth of a second—the sweep rate for conventional television displays. And systems capable of generating pictures requiring 100,000 polygons or more in minutes or even seconds are under development.

Hidden surfaces

The drive toward more realistic images has produced new algorithms capable of displaying more complex effects. At the same time the development of custom hardware is forcing some reexamination of well-established algorithms. One area of continuing debate is how best to find out which parts of an image are in front of other parts so that surfaces hidden from an observer's view can be eliminated. At the core of hidden-surface algorithm problems is a sorting procedure. Straightforward sorting routines that would compare every polygon in an image to every other polygon become impossibly large as the number of polygons grows. So nearly all algorithms use coherence—the continuity of an image from pixel to pixel, line to line, or frame to frame—to reduce the required computation.



To build a forest. A single, programmed leaf can be scaled up or down in size and rotated to produce the randomly varying leaves of a tree. The tree can be similarly scaled and rotated to form a forest.

Robert Marshall

One general class of hidden-surface algorithms sorts in what is known as object, or model, space. This orders polygons or other geometric primitives by distance from the observer and sorts them with the full, three-dimensional precision of the original model. Another general class sorts in image or screen space, determining distance pixel by pixel, but only to the resolution limit of the display device. Object-space algorithms tend to take more computation time but they allow inclusion of more complicated visual effects. They can, for example, include the multiple reflections of light among objects within an image.

Each approach has advantages and advocates. It comes down to "a trade-off between physically realistic images and computaton time," says Frank Crow of Ohio State, whose research group uses object-space methods because of the intricate visual effects the researchers wish to achieve. Edwin Catmull of Lucasfilm, working under different constraints, has a different perspective. He notes that "there is a classic paper ['A Characterization of Ten Hidden-Surface Algorithms,' by Ivan Sutherland, Robert F. Sproull, and Robert A. Schumacker; *Computer Surveys*, March 1974] pointing out several criteria for judging the efficiency of hidden-surface algorithms. Most people forget that algorithm choice depends on the complexity of the picture; you make different choices at a thousand polygons and at a hundred thousand polygons."

Subtle effects

But complexity is not the only key to greater realism. The human eye uses color and subtleties of shading as light reflects from a surface to identify an object and its material. A demonstration of the importance of these effects came from early algorithms that treated reflected light as having the color of its source rather than of the surface from which it is reflected. The results tended to make everything look like it was made of plastic.

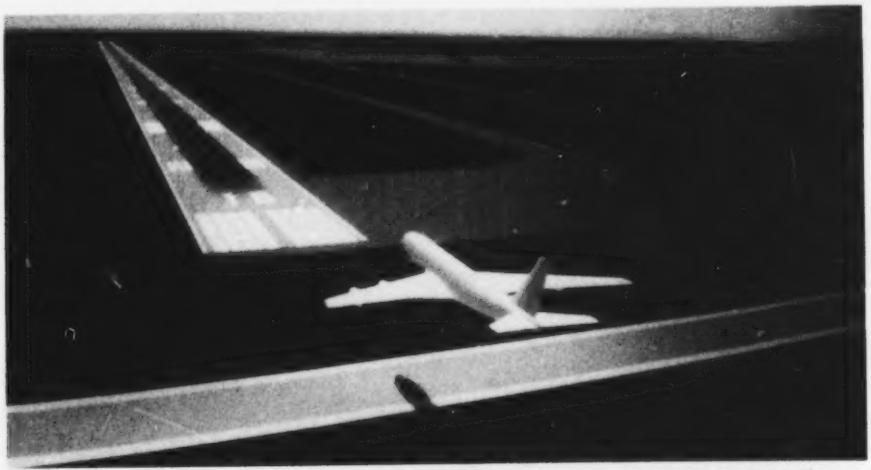
More recent algorithms treat diffuse reflections and highlights—specular reflections—differently. By varying the relative intensities of the two, a programmer can change the appearance of a surface from dull to glossy. More recent work at Cornell by Rob Cook, now at Lucasfilm, takes into account wavelength-by-wavelength changes in the spectral composition of reflected light as a function of the object's material makeup and its surface roughness. The result enables a programmer to display the difference between a metallic surface and a polished wood or plastic surface.

Still other work takes into account shadows, surface patterns and textures, and refraction and transmission of light. This yields, among others, algorithms that can display transparent objects. Most of these subtle effects require more than one pass through the display algorithms and take more computer time.

One of the most difficult problems in creating realistic images, and often a dead giveaway that a picture was created by a computer, is the effect called aliasing. It arises because pictures displayed on televisionlike screens exist as a series of pixels, each with a single color or intensity value, rather than as continuous images. Yet the object space sampled by a pixel may contain parts of two or more polygons, each with sharply different intensity values. The result is an image displayed with jagged lines or edges. Higher resolution displays reduce these ef-

---

Hammond is editor of Science 82.



Pilot training. A system stimulating events at an airport has a precomputed data base of about one million polygons. Objects modeled in several different levels of detail are automatically selected and shown to the pilot in sequences depicting real situations.

Evans and Sutherland

essors—in effect, many small computers—in a pipeline approach.

Going further

Henry Fuchs of the University of North Carolina has carried the idea of special-purpose hardware even further. The real bottleneck at present, he contends, is that a processor can get only one piece of data in or out of a memory chip at a given time. When data transmission is so rate limiting, the sophistication of fancy processors is wasted.

To get around this, Fuchs has analyzed computer-graphic problems from what he calls the memory point of view, and he has come up with radically new ways to describe objects. These new descriptions permit calculations by much simpler processors; in principle, it would be possible to build as many as one per pixel on just a few VLSI chips. It is not yet clear, Fuchs says, whether this approach will extend to enough display algorithms to make it practical.

Another custom approach to parallel processing is to encode computer graphics algorithms on VLSI chips—whose design, by a nice piece of circular happenstance, depends on computer graphic techniques. James Clark of Stanford (who credits Mark Hanna with the work), Turner Whitted of Bell Labs, and others are working on putting specific pieces of display algorithms in custom VLSI chips. The Stanford group has the distinction of developing the first practical application of such a chip to computer graphics—a graphics processor to generate and rotate polygons (see "Faster Switches, Smaller Wires, Larger Chips," *Mosaic*, Volume 13, Number 1). Clark and his colleagues are now working on what Clark calls "a custom smart image memory." It would divide a television screen into perhaps 64 regions, each with a processor, to create a system for drawing and manipulating graphic figures quickly and inexpensively.

The pressure for commercial applications is having an effect on computer graphic research. The largest (and, some observers say, the best) research groups in the field are now in industry. Many researchers feel, as Cornell's Donald Greenberg puts it, that "the frontier is the applications."

Despite this commercial appeal, computer graphics still faces many limitations in the kinds of images it can create. No one has tried to model people—George still has a long way to go—and even such common objects as the average well-filled bookcase would strain most systems. But these limitations may turn out to be unimportant, because the ability to combine computer graphic images with digitized photographs is already at hand.

Flight simulators include joint data of this kind, and people interested in applications in architecture are starting to combine models of houses or prospective urban developments with real pictures of their intended locations. As computer-generated pictures gain in realism, they can increasingly be combined with conventional photographs and images of all kinds. The end product will be hybrid images that, to the casual observer, will often not appear to have been generated by computers at all. •

The National Science Foundation contributes to the support of research discussed in this article through its Computer Systems Design Program.

If You Care About Science, You Will Want to Read

The National Science Foundation's

MOSAIC

NATIONAL SCIENCE FOUNDATION
MOSAIC
MOSAIC
Drilling into ocean sediment
Ice in the Sky

Discover that science can be timely, readable, *and* authentic.

"Scientists read *Mosaic* the way actors read *Variety*....I usually read it to find out about science in an area other than my specialty." That's the way one prominent oceanographer describes the usefulness of *Mosaic*. For that scientist, *Mosaic* is on target. His reason for reading the bimonthly magazine is precisely the reason it is published: to provide a perspective on the range of research areas on which the National Science Foundation's interests touch.

Mosaic is written and edited by experienced science journalists and authenticated by the researchers whose work it discusses. It can be of value to scientists who care about research areas other than their own, and to nonscientists who just care about science.

To subscribe to *Mosaic*, to send it as a serious gift to a serious friend, or to order single copies, use the order cards bound into this magazine.

National Science Foundation
Washington, D.C. 20550

Official Business
PENALTY FOR PRIVATE USE, $300

Postage and Fees Paid
National Science Foundation

U.S.MAIL

BULK RATE
U.S. POSTAGE PAID
PERMIT NO. G69
WASHINGTON, D.C.

University Research Equipment Expenditures Above $350 Million in 1980

Universities and colleges spent between $350 million and $400 million for separately budgeted research equipment in fiscal year 1980. This represented approximately 6 percent of all their separately budgeted research and development expenditures. This proportion ranged from 8 percent of the total academic R/D expenditures in the physical sciences to 4 percent in the social sciences. Federal funds accounted for two-thirds of these 1980 research equipment purchases, although the federal share differed significantly by science and engineering field.[1]

A recent NSF-sponsored study of four science and engineering subfields (organic chemistry, cell biology, solid-state physics, and electrical engineering) in 38 institutions produced the following data on the stock, condition, and use of research equipment in these subfields as of fall 1981.[2] The data pertain to items costing at least $5,000.

About one-fourth of research equipment in the four subfields combined was at least ten years old, a fourth was five to nine years old, and half was less than five years old. This pattern did not change significantly among the subfields except for electrical engineering; there, two-thirds of the research instruments were less than five years old.

Most institutions in the study had common facilities or pools containing significant amounts of research equipment. Among the larger research performers surveyed, 87 percent of the departments reported such facilities. The more expensive items tended to be located in these pools, where sharing among investigators could be maximized. Major items—research instruments costing at least $50,000, such as mass spectrometers and electron microscopes—accounted for more than 70 percent of the aggregate value of equipment items with a purchase cost of at least $5,000. Items in shared facilities represented 60 percent of the cost of all reported instruments.

Department heads, in describing the condition of equipment in large central pools, reported that about 10 to 12 percent of the items were in "poor" condition; 20 to 25 percent were in "fair" condition; 30 to 40 percent were in "good" condition; and 28 to 33 percent were in "excellent" condition.

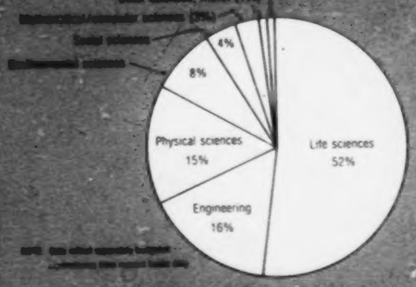
In all four subfields sampled, research equipment located in investigators' own laboratories was highly shared. In those laboratories, items costing at least $50,000 tended to have more users (seven to ten) than the average reported for all pieces of equipment (four to six). However, because of the greater demand for time on more expensive equipment, investigators reported a higher level of difficulty in gaining access to such items.

More than three-fifths of the research equipment purchases in the sample were wholly financed with federal funds, one-fifth solely with institutional funds, and one-tenth from other single sources. In addition, about one-tenth of the purchases were funded by a combination of funding sources; joint funding accounted for about 20 percent of the purchases in the $50,000-or-more category.

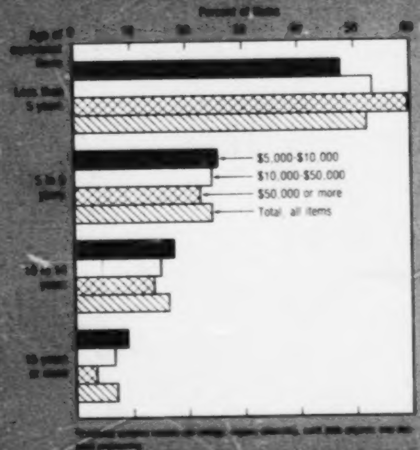
Five-sixths of research instruments in investigators' laboratories experienced less than 20 percent downtime from equipment failure. About two-thirds of investigators in the four subfields spent less than 5 percent of their research budget on maintenance and repair; one-fifth spent 5 to 10 percent, and one-tenth spent 10 percent or more, with available maintenance and repair resources tending to be applied to the more expensive items.

About 7 percent of all items were locally assembled. Researchers most often indicated that they had built their own equipment because no commercial equivalent existed.

Research equipment expenditures at universities and colleges by field: FY 1980



Age distribution of research equipment items in sampled departments as of 1981 by cost category[1]



---

[1] Data are derived from NSF's *Survey of Scientific and Engineering Expenditures at Universities and Colleges, FY 1980*. The data represent equipment purchases from current funds only and do not include research equipment purchases from plant (capital) funds, as in the case of newly constructed laboratories stocked with research instruments.

[2] NSF will do a full-scale survey during FY 1983 to acquire equipment data that are representative of the entire academic sector. The survey will cover engineering and computer and physical sciences; other fields will be surveyed during FY 1984.